SonUIT eRisk2025: Enhanced Depression Detection on Social Media via Filtering and Re-Ranking

Notebook for eRisk Lab at CLEF 2025

Nguyen Minh Son[1,2,*], Dang Van Thin[1,2]

[1]University of Information Technology-VNUHCM, Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

This paper presents our approach for eRisk 2025 Task 1: Search for Symptoms of Depression. The task involves ranking sentences from user writings based on their relevance to the 21 depression symptoms defined in the Beck Depression Inventory (BDI). A sentence is considered relevant if it provides information about the user's condition regarding a specific symptom. Our method follows a two-stage pipeline. In the first stage, we generate symptom embeddings from labeled training data and compute cosine similarity to rank sentences by semantic relevance. In the second stage, we apply re-ranking candidates using various strategies, including cross-encoders, BM25, or using larger embedding models. Our approach demonstrates strong performance, consistently ranking among the top three teams across all evaluation metrics out of a total of 17 participating teams.

Keywords

Depression Detection, BM25, Cross-Encoders, Re-Ranking, Sentence Transformers, BDI-II, Mental Health

1. Introduction

The CLEF eRisk 2025 lab [1] continues to advance the field of early risk detection on the Internet, with a strong focus on mental health applications such as depression detection. This year's edition includes three key tasks: Task 1: Search for Symptoms of Depression, which continues the 2024 version by ranking sentences based on their relevance to specific symptoms from the Beck Depression Inventory (BDI-II); Task 2: Contextualized Early Detection of Depression, a new task that analyzes full conversational threads to detect early signs of depression, and Pilot Task on Conversational Depression Detection via LLMs, which challenges participants to assess mental health states through interaction with language model personas.

In this work, we concentrate on Task 1, where the objective is to retrieve and rank sentences from user writings according to their relevance to 21 BDI-II depression symptoms. We propose a system that combines multiple filtering and reranking strategies to identify the top 1,000 candidate sentences per symptom. Our approach aims to capture both explicit and subtle indicators of mental health status, whether indicative of the presence or absence of symptoms. We discuss the performance of our system, highlight key challenges in re-ranking, and provide insights into future improvements in sentence-level mental health detection.

2. Related Work

The task was first introduced in the eRisk 2023 [2] Task 1 (Search for Symptoms of Depression), which required participants to rank sentences according to their relevance to each of the 21 items from the Beck Depression Inventory-II (BDI-II) [3]. The top-performing team, Formula-ML [4], utilized Transformer-based embeddings combined with soft cosine similarity over BDI-related terms to rank

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
22521254@gm.uit.edu.vn (N. M. Son); thindv@uit.edu.vn (D. V. Thin)
https://nlp.uit.edu.vn/ (D. V. Thin)
0000-0001-8340-1405 (D. V. Thin)

sentences effectively. In contrast, the BLUE [5] team generated synthetic training examples for each symptom using GPT and subsequently employed a dual transformer architecture for sentence ranking.

In the 2024 edition of the task, a total of 29 system runs were submitted by nine participating teams. The REBECCA team [6] applied Sentence Transformers for initial semantic encoding, filtered candidates based on cosine similarity, and performed re-ranking using GPT-4. The SINAI group [7] fine-tuned DistilRoBERTa on symptom relevance annotations and also explored a prompt-based variant using GPT-3 to assess the performance gap between supervised fine-tuning and prompting. DS@GT team [8]framed symptom detection as a supervised classification task but found that their classifiers were poorly calibrated for ranking metrics. APB-UC3M team [9] implemented an ensemble method that combined semantic similarity pipelines with a RoBERTa-based classifier to benchmark ensemble performance against single-model baselines. Finally, NUS-IDS team [10] fine-tuned multiple sentence-transformer models via contrastive learning, incorporating both BDI symptoms and Early Maladaptive Schemas and ensembled these predictors, demonstrating competitive performance across all major evaluation metrics.

3. Methodology

Building upon insights provided by the previous year's team, we adopted a refined approach informed by their empirical observations. Two key considerations guided the development of our method: (1) efficient preprocessing to reduce the dimensionality and noise within the large-scale dataset, and (2) re-framing the task not as a traditional supervised multi-label classification problem, but rather as a sentence-level semantic similarity task, which proved more effective for capturing contextual relevance.

3.1. Data Preprocessing

To standardize and clean the textual data, we first normalized all text by converting it to lowercase and removing punctuation, special characters, and non-linguistic symbols. In order to retain only sentences relevant to the user's personal experiences with symptoms, we implemented a lexical filtering step based on first-person pronouns (e.g., "I", "me", "my", "myself", "we", "us", "our", "ourselves"). This heuristic ensured that only sentences likely to reflect personal symptom descriptions were retained for further analysis.

3.2. Data Preparation

The labeled datasets from the 2023 and 2024 annotation cycles were utilized as our primary training corpus. For each symptom class, we employed a majority-vote criterion to determine relevance labels at the sentence level. Sentences annotated as relevant by the majority of annotators were retained, resulting in a high-confidence set of training examples for each symptom category.

3.3. Filtering and Re-ranking System

Our approach consists of a two-stage pipeline involving candidate sentence filtering followed by re-ranking, with optional preprocessing of the input text. The overall system architecture is illustrated in Figure 1. Details of each stage are described in the following sections.

3.3.1. Filtering

The first stage of our retrieval pipeline involved filtering the dataset to identify candidate sentences that are semantically similar to each symptom. We employed the Sentence Transformer model `all-MiniLM-L6-v2`. This model was selected based on its previously demonstrated effectiveness in semantic similarity tasks within our domain.

To generate robust and representative embeddings for each symptom class, we first grouped the training data by symptom label (denoted by the `query` field). For each symptom, all associated sentences

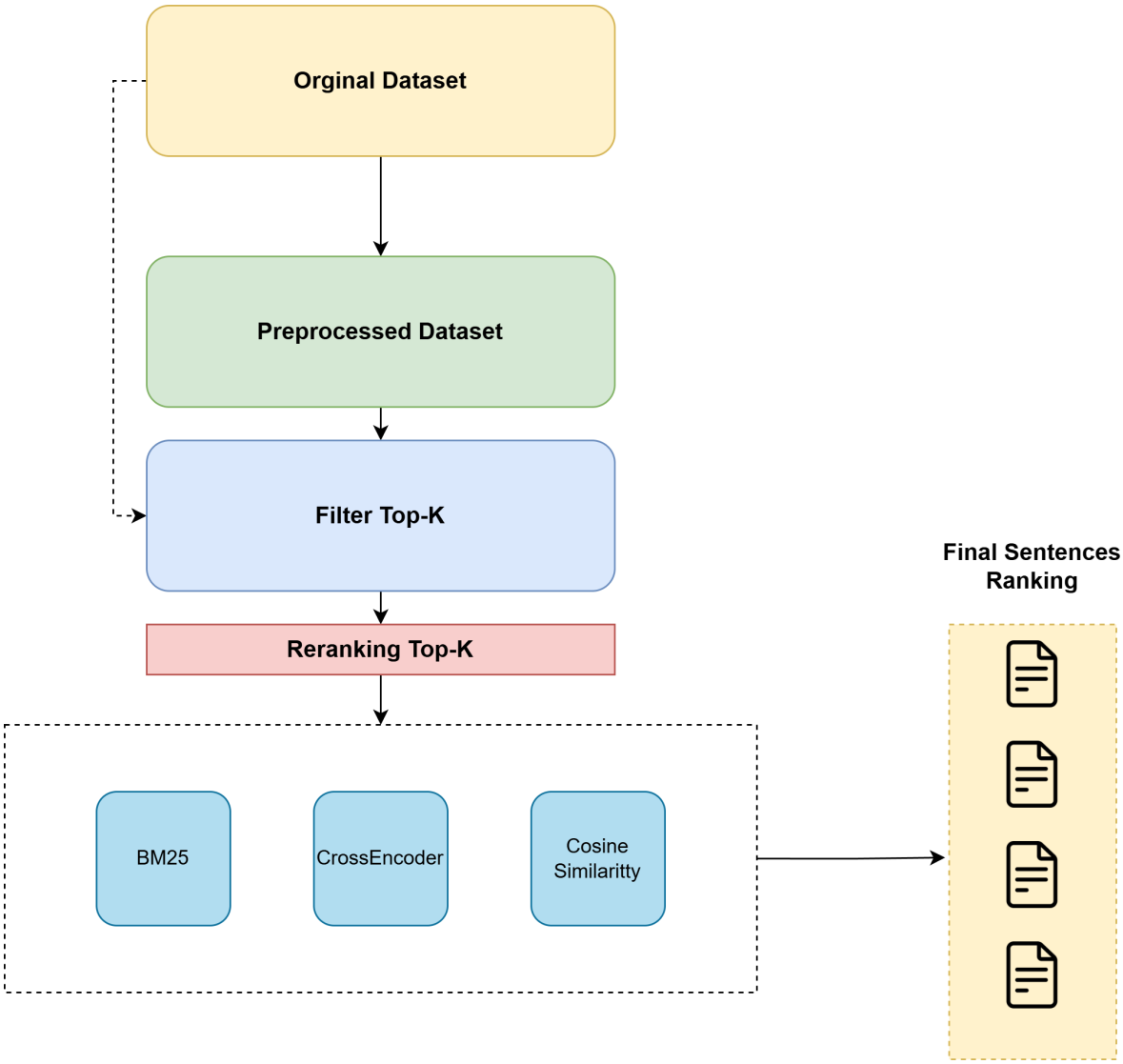

Figure 1: An overview of our system architecture.

were aggregated and encoded using a pre-trained transformer model. Each sentence was tokenized and processed in batches. The model's final hidden states (token embeddings) were extracted, and attention-weighted mean pooling was applied across valid tokens using the attention mask. This produced one vector per sentence. We then computed the average of these sentence-level vectors to obtain a single embedding representing the symptom. Our approach is inspired by the architecture of Sentence-BERT [11], which applies pooling over token embeddings to obtain sentence-level representations. While we do not fine-tune the model using contrastive loss, we similarly extract sentence embeddings from a Transformer model via pooling, and aggregate them to form class-level representations. The final process is as follows:

1. Group training data by symptom label (`query`).
2. Tokenize and encode all sentences for each symptom using `all-MiniLM-L6-v2`.
3. Extract token-level hidden states from the final layer.
4. Apply attention-weighted mean pooling across tokens to obtain one vector per sentence.
5. Average all sentence embeddings for each symptom to produce a final embedding.
6. Store these vectors as symptom-level representations.

With symptom embeddings in place, we calculate cosine similarity between each sentence in the

unlabeled corpus and the corresponding symptom embedding. The top k (where $k = 1000$) highest-scoring sentences for each symptom are selected as candidate sentences.

3.3.2. Re-ranking

To improve the quality of the final ranked list, we explore multiple re-ranking strategies applied to the filtered sentence candidates:

- **BM25**: A sparse retrieval method based on term frequency-inverse document frequency (TF-IDF) weighting, used to re-rank based on lexical similarity.
- **Cross-Encoder**: A transformer-based model that jointly encodes the query and candidate sentences to score them based on contextual relevance.
- **Cosine Similarity:** Sentence embeddings for candidates are re-computed using larger embedding models such as `BAAI/bge-large-en-v` [12] or OpenAI's `text-embedding-3-large`. For each candidate, cosine similarity is computed against all individual sentence embeddings associated with the target symptom. The final similarity score is obtained by averaging these values, resulting in a mean similarity score that captures overall alignment with the symptom representation. Candidates are then re-ranked based on this aggregated score.

Each re-ranking method was tested independently to isolate its impact on ranking performance. This modular design allows for direct comparison across different retrieval paradigms and embedding architectures.

4. Experimental Setup

We conducted five experiments to evaluate the effectiveness of different data processing and ranking strategies. Details of each configuration are outlined below. Although our methodology initially considered the use of cross-encoders for re-ranking, we excluded them from the final experiments due to their high computational cost and suboptimal performance when evaluated on the previous year's data, as shown in Table 1.

Based on observations from the previous year's data, the top 100 candidates retrieved during the initial filtering stage already exhibit high relevance. Therefore, in our approach, we focus re-ranking efforts primarily on the remaining candidates beyond this top set. It also revealed in Table 1 that aggressive preprocessing, such as removing punctuation and limiting sentences to those with first-person pronouns, may inadvertently exclude sentences that are still semantically relevant. As a result, using raw (non-preprocessed) data yielded higher performance in the 2024 dataset.

The five experimental configurations are as follows:

- **Experiment 1: Raw data + Filter**
  This baseline experiment uses the unprocessed dataset and applies sentence-level filtering based on semantic similarity. We employ the `all-MiniLM-L6-v2` Sentence Transformer to compute cosine similarity.
- **Experiment 2: Preprocessed Data + Filter**
  This setup mirrors Experiment 1 but applies preprocessing steps before filtering. These include converting text to lowercase, removing punctuation and special characters, and restricting the dataset to sentences containing first-person pronouns. The same filtering strategy using `all-MiniLM-L6-v2` is applied.
- **Experiment 3: Raw Data + Filter + Re-ranking (Cosine Similarity - Open Source Embedding Model)**
  Building upon Experiment 1, this experiment adds a semantic re-ranking step using the `BAAI/bge-large-en-v1.5` embedding model [12].

Table 1
Overall Performance Analysis on the 2024 Dataset

Method	MAP	R-PREC	P@10	NDCG@1000
preprocessed + mean-embedding	0.285	0.338	0.99	0.516
raw + attention-pooled	0.351	0.397	0.948	0.585
raw + mean-embedding	0.343	0.383	0.986	0.580
raw + mean-embedding + bm25	0.321	0.350	0.986	0.572
raw + mean-embedding + cross encoder	0.338	0.376	0.986	0.579
raw + mean-embedding + text-embedding-3-large	**0.371**	**0.427**	**0.986**	**0.590**
raw + mean-embedding + bge-large-en-v1.5	0.357	0.410	0.986	0.585

- **Experiment 4: Raw Data + Filter + Re-ranking (Cosine Similarity - OpenAI Embedding Model)** This experiment replaces the embedding model in Experiment 3 with OpenAI's `text-embedding-3-large`, a more robust embedding model base on MTEB.
- **Experiment 5: Raw Data + Filter + Re-ranking (BM25)** In this final experiment, re-ranking is performed using the BM25 algorithm, a classic sparse-retrieval method that scores candidate sentences based on lexical similarity to the symptom query.

Each experimental configuration, labeled config1 through config5, was evaluated using the official scoring metrics provided by the shared task, allowing for a systematic comparison across different retrieval and ranking strategies.

5. Results and Discussion

A comprehensive overview of all participating teams' performance, along with detailed explanations of the evaluation metrics, can be found in [1, 13].

In Task 1, we focused exclusively on majority voting, as our training data was constructed using majority rather than unanimity labels. The official results, shown in Table 2, demonstrate that our approach is highly effective, consistently ranking among the top three teams across all evaluation metrics. Although our system did not achieve the highest score in any single metric, it maintained strong and balanced performance overall.

Interestingly, unlike in the 2024 dataset, this year's results indicate that preprocessing had a positive impact on performance. Specifically, **config2**, the only configuration using preprocessed data, achieved the highest scores in three out of four metrics (MAP, P@10, and NDCG@1000). This suggests that, for the current dataset, preprocessing steps such as punctuation removal and sentence filtering may enhance relevance estimation.

Among our configurations, **config5**, which used BM25 for re-ranking, performed the worst, highlighting the limitations of traditional retrieval methods in this setting. In contrast, re-ranking with more powerful embedding models, such as **text-embedding-3-large** used in **config4**, yielded significant performance improvements over the baseline (**config1**).

Overall, the top-performing team in the shared task was INESC-ID, which achieved the highest scores across all four metrics using their unanimity-based configuration. The UET-Psyche-Warriors team, which employed a machine learning-based approach, also achieved high scores across all metrics. This is particularly noteworthy given that, in last year's evaluation, framing the task as a classification problem resulted in significantly lower performance.

6. Conclusion and Future Work

Our system demonstrates strong effectiveness in detecting early signs of depression by ranking user writings according to their relevance to specific depression symptoms. Despite these promising results,

Table 2
Task 1 (majority voting) Evaluation Among Best Runs of Top 3 Teams

Team	Run	MAP	R-PREC	P@10	NDCG@1000
SonUIT	config1	0.283	0.351	0.767	0.562
SonUIT	**config2**	**0.334**	**0.392**	**0.790**	**0.613**
SonUIT	config3	0.311	0.395	0.767	0.572
SonUIT	config4	0.328	**0.426**	0.767	0.578
SonUIT	config5	0.26	0.304	0.767	0.552
UET-Psyche-Warriors	5 machine learning	0.339	0.394	0.776	0.623
INESC-ID	max	0.350	0.407	0.648	**0.653**
INESC-ID	unanimity	**0.354**	**0.433**	**0.876**	0.575

there remains room for improvement, particularly in preprocessing strategies, as well as in the filtering and re-ranking methods employed.

Although our system did not achieve the highest score on any individual metric, its consistently balanced performance underscores the robustness of our approach. For future work, we plan to explore more sophisticated ensemble techniques and investigate the integration of cross-encoder models while addressing their computational costs. Additionally, we aim to experiment with machine learning approaches, inspired by the encouraging outcomes from other teams.

Moreover, further research into data preparation techniques, such as data augmentation and contrastive learning methods, could provide further performance improvements. These directions emphasize the need for ongoing investigation in this area to determine the optimal solutions for the early detection of depression symptoms.

Acknowledgments

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

Declaration on Generative AI

During the preparation of this work, we used GPT-4 and Grammarly in order to: check grammer, spelling, and edit the content for clarity and coherence. After using these tools, we reviewed and edited the content as needed and take full responsibility for the publication's content.

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] 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.

[3] 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. doi:10.1001/archpsyc.1961.01710120031004.

[4] 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.
[5] A.-M. Bucur, Utilizing chatgpt generated data to retrieve depression symptoms from social media, arXiv preprint arXiv:2307.02313 (2023).
[6] A. Barachanou, F. Tsalakanidou, S. Papadopoulos, Rebecca at erisk 2024: search for symptoms of depression using sentence embeddings and prompt-based filtering, Working Notes of CLEF (2024) 9–12.
[7] A. M. Mármol-Romero, A. Moreno-Muñoz, P. Álvarez-Ojeda, K. M. Valencia-Segura, E. Martínez-Cámara, M. García-Vega, A. Montejo-Ráez, Sinai at erisk@ clef 2024: Approaching the search for symptoms of depression and early detection of anorexia signs using natural language processing, Working Notes of CLEF (2024) 9–12.
[8] D. Guecha, A. Potdar, A. Miyaguchi, Ds@ gt erisk 2024: Sentence transformers for social media risk assessment, arXiv preprint arXiv:2407.08008 (2024).
[9] A. Bascuñana, I. S. Bedmar, Apb-uc3m at erisk 2024: natural language processing and deep learning for the early detection of mental disorders, Working Notes of CLEF (2024) 9–12.
[10] B. H. Ang, S. D. Gollapalli, S.-K. Ng, Nus-ids@ erisk2024: ranking sentences for depression symptoms using early maladaptive schemas and ensembles, Working Notes of CLEF (2024) 9–12.
[11] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084. arXiv:1908.10084.
[12] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.
[13] 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.