BDI-batch: Leveraging Standardized Clinical Questionnaires for Contrastive Learning in Psychological Marker Retrieval

Marcos Fernández-Pichel and David E. Losada
Centro Singular de Investigación en Tecnoloxias Intelixentes (CiTIUS),
Universidade de Santiago de Compostela (USC), Spain
{marcosfernandez.pichel, david.losada}@usc.es

Abstract

Automatic detection of signs of depression has received increasing attention in recent years, evolving from user-level classification (binary) to fine-grained symptom detection. Despite recent advances, state-of-the-art sentence retrieval models still struggle to discriminate between sentences associated with closely related symptoms. In this work, we propose `BDI-batch`, a novel contrastive learning method that leverages clinically grounded sentence pairs to encode fine-grained distinctions among depression-related symptoms. The method exploits a standardized clinical questionnaire and does not rely on annotated data. We demonstrate the effectiveness of our approach on two early risk detection datasets and under multiple retrieval strategies. Additionally, we explore the use of LLMs to generate synthetic contrastive samples and observe that their performance falls short of questionnaire-derived pairs. Finally, we conduct an error analysis that reveals the most common errors of the sentence retrievers.

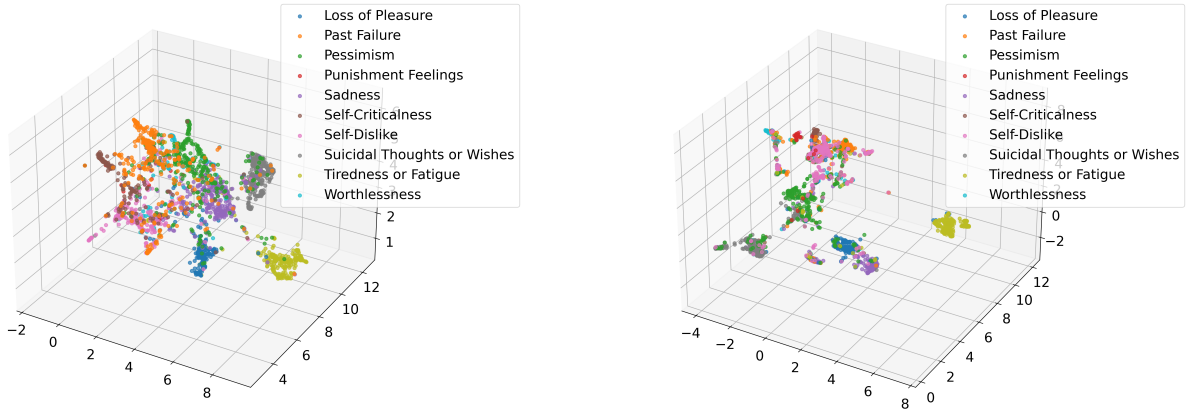
1 Introduction

Depression is a prevalent mental disorder characterized by a range of cognitive, physical, and emotional symptoms that significantly impair daily functioning. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (APA, 2022) defines a set of symptoms associated with this disorder, including persistent sadness, fatigue, or changes in sleep patterns. According to the WHO (WHO, 2025), around 4% of the global population experiences depression, including 5.7% of adults (4.6% among men and 6.9% among women). Despite its prevalence, diagnosis and treatment remain limited: in high-income countries, only about one-third of affected individuals receive appropriate care (Evans-Lacko et al., 2018).

In recent years, social media has emerged as a valuable source of insight into users' mental states. As individuals increasingly disclose thoughts and emotions related to their mental wellbeing online (Ríssola et al., 2021; Crestani et al., 2022), computational methods have emerged as a way to complement clinical assessment and deepen our understanding of the disorder. Early approaches addressed binary classification challenges (e.g., depressed vs. non-depressed), whereas more recent studies have focused on identifying specific symptoms in online text (Pérez et al., 2022; Zhang et al., 2022b; Nguyen et al., 2022; Zhang et al., 2022a; Pérez et al., 2023).

The CLEF eRisk shared-data task (Losada et al., 2019) has been a central venue for advancing research in this area. Recent editions of eRisk introduced a symptom retrieval task, where a large corpus of sentences was given to the participants and the goal was to retrieve sentences relevant to each of the symptoms in the Beck Depression Inventory (BDI) (Pérez et al., 2025). This task is particularly challenging because of the blurred distinction between some symptoms. For example, the BDI items 'Sadness' and 'Crying' exhibit substantial overlap, and a sentence retriever may struggle to identify textual excerpts that are truly aligned with each of these items. Even state-of-the-art sentence embeddings face difficulties in distinguishing between similar BDI items (Pérez et al., 2025). For instance, in Figure 1 we can observe that the `all-mpnet-base-v2` model (left-side plot) struggles to separate several depression-related symptoms (e.g., "Self-dislike", "Past-failure", "Pessimisim", and "Self-criticalness"), whereas our model (right-side plot) makes a clearer separation among these groups (without degrading already well-clustered symptoms such as "Tiredness or Fatigue").

Contrastive learning has recently shown promise in improving semantic discrimination among similar sentences (Izacard et al., 2021; Gao et al., 2021). Particularly, the inclusion of hard negatives encour-



(a) Base

(b) BDI fine-tuned

Figure 1: Sentences annotated as relevant to ten BDI symptoms. The left-side graph plots the UMAP representations of their `all-mpnet-base-v2` embeddings while the right-side graph plots the UMAP representations of our BDI-based variant.

ages models to distinguish between thematically related but semantically distinct samples (Jiang et al., 2024; Kim et al., 2024; Liu et al., 2024). Inspired by these advances, we explore contrastive learning to enhance sentence representations for depression symptom retrieval. Specifically:

- We propose a novel in-batch contrastive learning method, `BDI-batch`, for generating effective embeddings tailored to depression symptom retrieval.
- We introduce a strategy for automatically obtaining negative examples. This method does not require labeled data and leverages a standardized screening questionnaire to generate contrastive pairs of samples. We demonstrate the effectiveness of this method against several baselines, including state-of-the-art retrieval models.
- We evaluate the use of open and proprietary Large Language Models (LLMs) for generating synthetic contrastive samples.
- We provide a qualitative analysis of the most frequent errors observed in sentence embedding models.

2 Related Work

2.1 Automatic Detection of Depressive Signals

Previous research on the automatic detection of signs of mental health disorders has relied on Convolutional Neural Networks (CNNs) (Yates et al., 2017; Rao et al., 2020), Recurrent Neural Networks (RNNs) (Trotzek et al., 2018; Skaik and Inkpen, 2020), and, more recently, Transformer-based architectures (Martínez-Castaño et al., 2020; Bucur et al., 2023, 2021; Poświata and Perełkiewicz, 2022). Early approaches addressed this task as a binary classification problem, oriented to distinguishing between depressive and non-depressive users. While effective in controlled settings, this formulation limits the applicability of such models in real clinical contexts, where symptom-level understanding is often required.

More recent studies have explored ways to incorporate clinical interpretability into computational models. For example, Nguyen et al. (2022) integrated PHQ-9 symptom dimensions into their predictions, achieving interpretability comparable to medical questionnaires. Similarly, Wang et al. (2024) proposed a two-stage approach that first retrieves sentences deemed to be relevant to Beck Depression Inventory symptoms and then estimates the overall level of depression. In line with these developments, the CLEF eRisk 2023 shared task introduced a new challenge focused on symptom retrieval (T1 task) (Parapar et al., 2023), thus framing depression detection as a sentence-level retrieval problem. Building on this task, Pérez et al. (2025) released the DepreSym dataset, which contains approximately 21.5K sentences annotated according to their relevance to specific BDI symptoms. This dataset, obtained by pooling rankings from 37 participating search systems, is one of the collections used in our experiments.

In recent years, commercial mental health applications have also gained traction. These include

Limbic Access (Habicht et al., 2024), which supports clinical triage within the UK's NHS; Woebot (Fitzpatrick et al., 2017), which delivers conversational Cognitive Behavioral Therapy (CBT); and Kintsugi Voice (Mazur et al., 2025), which enables acoustic-based depression screening. While these systems support clinical screening, triage, or intervention, our work addresses fine-grained symptom retrieval through clinically grounded representation learning.

2.2 Contrastive Learning

This paper is also related to a second strand of literature on contrastive learning for sentence embeddings. In this area, models such as SimCSE (Gao et al., 2021), Contriever (Izacard et al., 2021), and HN-CSE (Liu et al., 2024) have demonstrated that the inclusion of hard negatives (semantically similar but meaningfully distinct samples) can significantly improve representation quality for text retrieval and semantic matching. Our work builds upon these previous studies by introducing a novel method for the automatic generation and selection of hard negatives, tailored to the task of retrieving relevant sentences for specific depression symptoms. In an innovative manner, we utilize a standard medical questionnaire to generate these challenging negative samples, and, similar to (Bucur, 2023), we also evaluate the ability of LLMs to produce effective synthetic data. Note that we do not contribute a new contrastive objective or architecture. Instead, our contribution lies in deriving clinically grounded hard negatives from the structured response options of a validated questionnaire.

3 Beck Depression Inventory

The Beck Depression Inventory is a validated clinical questionnaire designed to measure 21 depressive symptoms, spanning emotional (e.g., *"Pessimism" or "Sadness"*), cognitive (e.g., *"Indecision"*), and physical (e.g., *"Fatigue"*) dimensions (Beck et al., 1996a,b). For each symptom, the questionnaire provides four possible responses (e.g., for "*Sadness*" the choices are "*I do not feel sad*", "*I feel sad much of the time*", "*I am sad all the time*", and "*I am so sad or unhappy that I can't stand it*"). It is a psychometric instrument designed to rapidly screen for and measure the severity of depressive symptoms. Clinicians primarily use it to establish a baseline score and monitor a patient's progress and treatment effectiveness over time.

Under CLEF eRisk, the BDI has been leveraged to stimulate research on retrieving social media excerpts that may signal depressive symptoms (Losada et al., 2019, 2020; Ríssola et al., 2021; Crestani et al., 2022; Parapar et al., 2021, 2023, 2024, 2025). Specifically, eRisk provided a large collection of sentences published by social media users and the goal was to retrieve sentences that are deemed as relevant for each BDI symptom.[1] This challenge has been addressed using state-of-the-art retrieval technology, including dense retrievers with a dual-encoder architecture. However, as shown in Fig. 1, encoded representations of sentences from different BDI symptoms tend to overlap. General-purpose sentence embedding models, such as `all-mpnet-base-v2`[2] (Song et al., 2020) or Contriever[3], were trained on massive text corpora drawn from highly diverse sources, but still fail to capture the fine-grained distinctions between different depressive symptoms.

4 BDI-batch

Given a BDI item (symptom) and a sentence posted online, a natural approach is to encode both with a sentence embedding model and estimate relevance based on the similarity between the encoded representations. Indeed, Pérez et al. (2025) reported that a dual encoder based on `all-mpnet-base-v2` achieved the strongest performance in retrieving sentences relevant to BDI symptoms. This sentence-transformers model derives from a pretrained mpnet-base model that was subsequently fine-tuned in a one billion sentence pairs dataset. By injecting massive numbers of contrastive pairs from multiple natural language datasets (Bowman et al., 2015) (e.g., question-answer pairs, duplicate question pairs, entailment pairs, and so forth), the model was taught to capture broad patterns of semantic similarity. Similarly, Contriever (Izacard et al., 2021) is a retrieval model trained with the Momentum Contrast (MoCo) objective (He et al., 2020), enabling it to learn from a broad and diverse set of negative examples. Although valuable as general-purpose semantic search devices, we argue that these methods lack the level of detail required

[1] This sentence retrieval task was conducted in 2023, 2024 and 2025 as eRisk Task 1 (T1)

[2] `all-mpnet-base-v2`: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[3] Its adapted version to sentence-level representations: https://huggingface.co/nishimoto/contriever-sentencetransformer

Pair	Symptom
I feel sad much of the time — I do not feel sad	Sadness
I cry more than I used to — I feel like crying, but I can't	Crying
I am not discouraged about my future — I do not expect things to work out for me	Pessimism
I don't feel I am being punished — I expect to be punished	Punishment Feelings
I dislike myself — I am disappointed in myself	Self-dislike
I don't enjoy things as much as I used to — I can't get any pleasure from the things I used to enjoy	Loss of Pleasure

Table 1: in-batch contrastive examples for the BDI item Sadness (5 most similar symptoms included in the batch).

to discriminate between specific symptoms linked to depression.

In standard sentence embedding models, no informed selection of negative pairs was applied during training and, thus, sentences associated with topically related symptoms such as "*Sadness*" and "*Crying*" end up close in the embedding space. We need to go beyond this basic representational approach by instructing the models to identify which sentences are indeed far. To that end, we propose `BDI-batch`, a strategy that leverages the responses of the BDI questionnaire to generate contrastive pairs. To obtain pairs of sentences linked to a given BDI item, we simply sample two random responses for the BDI item. For example, for "*Sadness*", a pair would be {"*I feel sad much of the time*","*I am so sad or unhappy that I can't stand it*"}. This gives up to $Permutations(4, 2)$ unique pairs derived for each BDI item.[4]

Now, if we randomly choose one pair from each BDI item and form a batch of 21 pairs, then we can apply Multi-Negative Ranking Loss,[5] which encourages each pair of sentences to be pulled together while pushing it away from all other pairs in the batch. This has the effect of bringing closer together the sentences associated with a given BDI symptom while simultaneously pushing apart those linked to other items. This process can be repeated multiple times, with each iteration selecting random pairs of responses that represent each BDI item.

However, the aforementioned approach may produce many uninformative contrastive pairs, given the semantic gap that exists among several BDI items. We therefore opt to group the BDI items into clusters of related symptoms and create smaller batches containing BDI items that are somehow similar. In this way, our method forces the model to learn fine-grained distinctions (i.e., induces harder contrastive pairs). To decide which BDI items should be included in a batch, we first encode all items[6] and compute pairwise similarities. For each BDI item, its k most similar symptoms are included into its batch (see Table 1). The goal is to force the encoder to separate nearby symptoms in the embedding space, thereby improving its ability to discriminate subtle symptom distinctions in the downstream retrieval task. Note that we always have 21 batches of size k (one for each BDI item).

5 Experiments

Following the evaluation protocol proposed by Pérez et al. (2025), we treat each BDI item (depression symptom) as an independent search case. For generating queries against the corpus, we leverage each possible response from the questionnaire for the target item. The search performance for each symptom is obtained by averaging the retrieval effectiveness metrics across these queries. This allows us to remain fully comparable with the reference baselines while providing a stable estimate of sentence-level retrieval performance.

According to the relevance criteria (Pérez et al., 2025), a sentence is considered relevant to a given symptom "*if it is on-topic but also provides explicit information about the individual's state related to the symptom*". This subtle notion of relevance extends beyond topicality, and indeed, we observed that the models (e.g., `all-mpnet-base-v2`) had a tendency to retrieve sentences that were on-topic but did not accurately describe the person's own psychological state. We therefore incorporated a simple method based on part-of-speech (POS) tagging to detect first-person constructions[7] and promoted on-topic sentences that fulfill this criterion.[8]

[4] A couple of BDI items have more than 4 possible responses and, therefore, these items allow us to generate more positive pairs.

[5] https://sbert.net/docs/package_reference/sentence_transformer/losses.html#multiplenegativesrankingloss

[6] We encoded the title and all the textual responses of each BDI item with the model `all-mpnet-base-v2`

[7] POS tagging was performed using SpaCy's *en_core_web_sm* model (Honnibal et al., 2020).

[8] Specifically, we ensured that all sentences containing first-person expressions were ranked above those that did not. To achieve this, we simply added a large constant offset to every

For training the BDI-batch embedding models, we set the batch size to $k = 10$ symptoms. The combinatorial space of possible batches is large, since each symptom yields at least 12 option pairs. To account for this, we need a number of epochs that is large enough to adequately cover this broad training example space, but at the same time small enough to avoid overfitting. We opted for selecting 10 epochs as a reasonable trade-off between exposure to diverse contrastive configurations and generalization. Both hyperparameters might have an important effect in performance, thus the sensitivity of our method to different epochs and batch sizes is studied in Section 6.3.

For our experiments, we used the **collections** from the eRisk T1 2023 and 2024 tasks (Pérez et al., 2025; Parapar et al., 2024). [9] These tasks consisted of identifying sentences relevant to each BDI symptom. Participants received a large collection of candidate sentences (extracted from social media) and were instructed to rank them by their relevance to each symptom. The eRisk organizers then applied a pooling strategy on the submitted runs and, subsequently, human assessors annotated each pooled sentence with a label that indicates its relevance to the target symptom.

We compare our BDI-batch approach against several **baselines**. BM25 is a classic term-based retrieval method, while BM25+CE[10] extends it by incorporating a cross-encoder re-ranking step (Thakur et al., 2021). We also experiment with several dense retrievers based on dual-encoder architectures, where BDI responses and candidate sentences are both encoded into dense representations. For instance, ANCE (Xiong et al., 2020) consists of a bi-encoder architecture that employs Approximate Nearest Neighbor (ANN) indexes to create robust training instances. Two additional key baselines are dense retrieval methods based on the embedding models all-mpnet-base-v2 and Contriever. For these methods, we evaluate both their unmodified form (i.e., using their original embeddings) and our variant, which performs the BDI-based embedding adaptation process outlined above. To ensure a fair comparison, the first-person promotion was applied to all baselines. Our objective is not to optimize an end-to-end retrieval system, but rather to generate embeddings that effectively represent the BDI symptoms. This improvement is potentially complementary to other advances in the retrieval process, and we leave for future work the integration of these representations into other retrieval strategies.

Following Pérez et al. (2025), retrieval effectiveness is measured using a recall-oriented metric, recall at 100 (R@100), and two NDCG-variants ($NDCG$@10 and $NDCG$@1000). The experiments were run in a single node with a 5090 GPU and 32GB of VRAM. Learning rate and warm-up steps were fixed to $5e-5$ and 100, respectively. Our code is publicly available for the research community to use.[11]

6 Results

In Table 2, we report the search results for both collections. We name our models with the prefix bdi followed by the name of the original model they were trained on. We conducted Wilcoxon signed-rank tests ($\alpha = .05$) to compare our dense retrieval variants with their original counterparts.

In both datasets, the BDI-batch variants significantly outperform the baselines and yield improvements over the original retrievers that are statistically significant. The bdi-all-mpnet-base model achieves the best results overall. This is a natural outcome, given that all-mpnet-base-v2 was already the best-performing baseline. Observe also that bdi-contriever is substantially better than contriever, also demonstrating the benefits of our embedding-creation process.

In the 2024 collection, the performance figures of the baseline methods were better. Particularly, ANCE achieved the best NDCG@10 performance among the baselines. However, our variants outperformed these results. Again, the Contriever-based variant experienced the largest increase. These results demonstrate the effectiveness of our contrastive learning approach in improving psychological marker retrieval.

Our novel way to leverage clinical questionnaire data to generate contrastive examples was highly effective. For instance, Figure 2 presents the symptoms exhibiting the largest relative improvements under our BDI-batch fine-tuning method, as well as those showing the smallest gains (or even performance declines). Only two symptoms are adversely

sentence that met this criterion. In Appendix A we further discuss the effect of this stage through an ablation study.

[9] Collections were obtained upon request to eRisk organizers.

[10] We used cross-encoder/ms-marco-MiniLM-L-6-v2 model for re-ranking

[11] https://github.com/MarcosFP97/bdi-batch

Dataset		Model	R@100	NDCG@10	NDCG@1000
eRisk T1 2023	Baselines	BM25	.141	.356	.404
		BM25 + CE	.141	.704	.446
		ANCE	.274	.802	.664
		all-mpnet-base-v2	.291	.834	.698
		contriever	.273	.761	.686
	Ours	bdi-all-mpnet-base-v2	***.372***$^{\uparrow}$	***.947***$^{\uparrow}$	***.817***$^{\uparrow}$
		bdi-contriever	*.324*$^{\uparrow}$	*.823*$^{\uparrow}$	*.743*$^{\uparrow}$
eRisk T1 2024	Baselines	BM25	.177	.628	.625
		BM25 + CE	.177	.858	.649
		ANCE	.277	.938	.837
		all-mpnet-base-v2	.268	.926	.805
		contriever	.265	.880	.819
	Ours	bdi-all-mpnet-base-v2	***.315***	***.976***$^{\uparrow}$	***.948***$^{\uparrow}$
		bdi-contriever	*.303*$^{\uparrow}$	*.941*	*.921*$^{\uparrow}$

Table 2: Retrieval performance results. We mark in italics if our `BDI-batch` strategy outperforms its base counterpart. In bold, the best performer for each collection and measure. The symbol ↑ marks the cases where our variants led to statistical significant improvements with respect to their original variant.

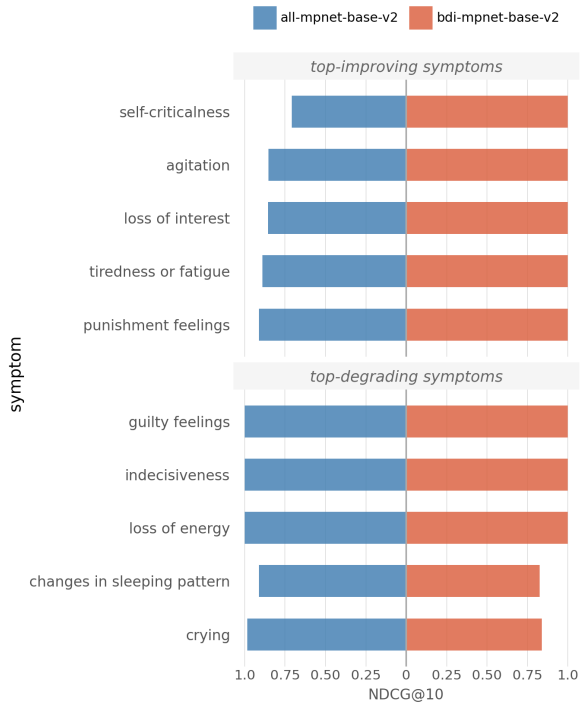


Figure 2: NDCG@10 performance of the original bi-encoder and our BDI fine-tuned model.

affected by our approach. As discussed earlier, this behavior can be attributed to the inherent difficulty of the task and to cases where symptoms exhibit substantial overlap, which can hinder precise discrimination. Overall, however, our method yields consistent performance improvements across most symptoms.

To better understand the impact of our approach, we visualize the embeddings of relevant sentences for both the original model and our `BDI-batch` encoder (see Appendix B). Our `BDI-batch` models are publicly available to use.[12]

6.1 Synthetic data generated by LLMs

As a natural alternative to train the embeddings, we evaluated LLMs' ability to generate synthetic pairs. We experimented with a proprietary model, GPT-4 (Achiam et al., 2023), and an open model, Qwen3-8B (Bai et al., 2023),[13] and instructed these models to generate synthetic sentences that are relevant to the BDI symptoms (see prompt in Appendix C, Figure 5). For each BDI item, we generated as many synthetic sentences as those available in the real clinical questionnaire and, following the method described in Section 4, the resulting sentences were leveraged to obtain contrastive pairs that were used to fine-tune the `all-mpnet-base-v2` model.

As shown in Table 3, fine-tuning with LLM-generated contrastive examples did not outperform the use of BDI-derived pairs. One possible explanation is that, despite explicit prompting, the models often failed to produce truly separate sentence pairs. For example, for the "Crying" symptom, GPT-4 generated "I express my emotions appropriately". In contrast, pairs obtained from the real questionnaire are inherently more distinctive of specific symptoms and therefore provide a clearer signal for guiding the contrastive learning process.

[12] https://huggingface.co/citiusLTL/bdi-all-mpnet-base-v2. Note that access to the models' weights does not grant or imply any license to use the official BDI-II.

[13] We used the Q4 quantized version: https://ollama.com/library/qwen3:latest

Dataset		Model	R@100	NDCG@10	NDCG@1000
eRisk T1 2023	BDI data	bdi-all-mpnet-base-v2	.372	.947	.817
	LLM data	gpt4-all-mpnet-base-v2	.332$^{\downarrow}$	.901$^{\downarrow}$	.784$^{\downarrow}$
		qwen-all-mpnet-base-v2	.301$^{\downarrow}$	.844$^{\downarrow}$	.724$^{\downarrow}$
eRisk T1 2024	BDI data	bdi-all-mpnet-base-v2	.315	.976	.948
	LLM data	gpt4-all-mpnet-base-v2	.296$^{\downarrow}$	.955	.914$^{\downarrow}$
		qwen-all-mpnet-base-v2	.287$^{\downarrow}$	.926$^{\downarrow}$	.868$^{\downarrow}$

Table 3: Contrastive learning with BDI-data vs LLM-generated data. The symbols ↑ and ↓ mark statistical significant improvements or deteriorations with respect to the BDI variant.

6.2 Comparison with eRisk's participants

To better contextualize the performance of our best models, we reproduced the similarity-based strategies used by the top-ranked eRisk systems in our sentence-pooled retrieval setting (one retrieval run per BDI option). For the 2024 collection, the AP3CM team (Bascuñana and Bedmar, 2024) obtained the highest precision scores using `all-MiniLM-L12`, a smaller but more effective model for sentence retrieval (Wang et al., 2020). Meanwhile, the NUS IDS team (Ang et al., 2024) achieved the best recall-oriented results by combining an ensemble of sentence encoders with multiple query formulations (referred to as exemplars). Their ensemble first retrieved top-k candidates independently with `all-mpnet-base-v2`, `all-MiniLM-L12`, and `all-MiniLM-L6`, merged the candidate sets, and then re-ranked them by averaging the exemplar-based similarity scores across models. In our setting, their exemplars can be naturally instantiated as the four BDI options for each symptom, allowing us to adapt their ensemble strategy to our setting.

Table 4 shows that our `BDI-batch` model outperforms both retrieval approaches. This result is particularly relevant because it demonstrates the robustness of our method when compared against more specialized and even ensemble systems. Notably, these approaches outperform all baseline models reported in Table 2, yet our model still achieves higher effectiveness.

6.3 Ablation study

To understand the effect of the number of epochs and number of symptoms in each batch, we experimented with the `bdi-all-mpnet-base-v2` model on the eRisk T1 2023 with varying numbers of these hyperparameters (see Table 5). Note that we also evaluated a configuration where pairs of sentences from any other symptom were included in the batches. Regarding epochs, both configurations (10 and 20) seem equally effective. As can be seen, increasing training epochs with respect to our default configuration did not provide further value, most likely due to excessive repetitions of the same pairs. Regarding the number of symptoms per batch, 5 symptoms and all symptoms seem suboptimal. The latter considers all remaining symptoms as potential sources of negative pairs and, therefore, fails to generate genuinely hard negatives. The slight underperformance of the 5-symptom configuration may stem from the implicit association among many BDI items, which makes it beneficial to contrast a given item with a wider range of related symptoms. Taken together, these results validate the design of `BDI-batch`, since exposing the model to sentences from related symptoms in the same batch promotes finer-grained differentiation in the embedding space.

7 Error Analysis

To further understand the benefits of this representational approach, we selected the two symptoms that showed the largest precision gains achieved by `BDI-batch` ("Agitation" and "Self-dislike") and performed a manual inspection of the top-5 retrieved sentences (see Table 6). The original encoder retrieves several irrelevant sentences for both symptoms, whereas our model achieves perfect P@5. For "Agitation", the baseline ranks topically related but irrelevant sentences (e.g., "I feel extremely jittery and kinda angry and I keep adjusting my clothes") as well as sentences that are relevant to closely related symptoms (e.g., "I am way more irritable, angrier, less patient, etc!", which corresponds to the "Irritability" symptom). For "Self-dislike", the baseline again retrieves sentences that are semantically close but do not explicitly express that the individual experiences the symptom (e.g., "I am disappointed").

By inspecting additional symptoms and errors made by the Contriever model, we observed that

	Model	R@100	NDCG@10	NDCG@1000
Ours	bdi-all-mpnet-base-v2	.315	.976	.948
Participants' approaches	AP3CM	.291$^{\downarrow}$	.941$^{\downarrow}$	.875$^{\downarrow}$
	NUS IDS	.294$^{\downarrow}$	.952	.851$^{\downarrow}$

Table 4: Comparison of our BDI contrastively fine-tuned model with the top eRisk 2024 systems, re-evaluated on our sentence-level retrieval setting. We also marked (↑↓) the statistical significant differences with respect to our model.

Hyperparams	R@100	NDCG@10	NDCG@1000
10 epochs & 5 symptoms	.360	.939	.801
10 epochs & 10 symptoms	.372	.947	.817
10 epochs & 15 symptoms	.367	.945	.818
10 epochs & all symptoms	.367	.940	.821
20 epochs & 5 symptoms	.360	.941	.796
20 epochs & 10 symptoms	.370	.948	.812
20 epochs & 15 symptoms	.364	.945	.813
20 epochs & all symptoms	.366	.944	.812

Table 5: Varying the number of epochs and batch sizes (`bdi-all-mpnet-base-v2` model, eRisk T1 2023).

Symptom	Model	Top-5 results
Agitation	all-mpnet-base-v2	"Its hard to explain, its like I feel restless and incapable of action at the same time." (Rel.)
		"I feel bursts of positivity...then restlessness and end up doing nothing." (Rel.)
		"I'm restless, can't sit still, can't concentrate, etc." (Rel.)
		"I feel extremely jittery and kinda angry and I keep adjusting my clothes." (Irrel.)
		"I am way more irritable, angrier, less patient, etc!." (Irrel.)
	bdi-all-mpnet-base-v2	"I'm restless, can't sit still, can't concentrate, etc." (Rel.)
		"Its hard to explain, its like I feel restless and incapable of action at the same time." (Rel.)
		"But right now I just feel restless and weird." (Rel.)
		"Ive tried meditation but I just become restless as hell" (Rel.)
		"I am also restless." (Rel.)
Self-dislike	all-mpnet-base-v2	"I am so disappointed in myself." (Rel.)
		"I'm dissapointed in myself." (Rel.)
		"I am disappointed in myself right now I edited" (Rel.)
		"I am disappointed" (Irrel.)
		"I am disappointed" (Irrel.)
	bdi-all-mpnet-base-v2	"I am so disappointed in myself." (Rel.)
		"i really hate myself" (Rel.)
		"I really hate myself" (Rel.)
		"I'm dissapointed in myself." (Rel.)
		"I'm just so disappointed in myself that it's not the case." (Rel.)

Table 6: Top-5 results for `all-mpnet-base-v2` model and our BDI fine-tuned model.

mistakes fall into the same two categories identified earlier: i) selecting a sentence that is relevant to a different but closely related symptom (e.g., for "Changes in Sleeping Patterns", the model retrieved sentences relevant to "Loss of Interest in Sex"), and ii) selecting a topically related sentence that does not clearly express the symptom (e.g., Contriever considered "Normally, when I go to sleep, I try to sleep as fast as possible" relevant to "Changes in Sleeping Patterns"). The first type of error accounts for approximately 45% of cases, while the second is slightly more frequent.

8 Discussion

A central insight emerging from our study concerns the nature of hard negatives in the context of clinical symptom retrieval. Unlike many settings where negatives are clearly distinct, depressive symptoms are inherently overlapping. As a result, a hard negative is not merely a semantically similar but irrelevant instance, but rather a sentence that plausibly expresses a related symptom. The effectiveness of `BDI-batch` suggests that contrastive learning can be meaningfully adapted to this setting when negative examples are grounded in subtle clinical dis-

tinctions. We observed a trade-off between symptom granularity and semantic overlap. Our ablation results indicate that using too few symptoms per batch limits the diversity of contrasts, while including all symptoms dilutes the hardness of negatives by introducing pairs that are trivially separable. Intermediate batch sizes appear to strike a balance by exposing the model to a sufficiently broad yet still semantically challenging set of related symptoms. This finding reflects a more general limitation of representation learning in highly correlated label spaces, where excessive granularity may not translate into better discrimination unless supported by carefully structured training signals.

Finally, the comparison between questionnaire-derived contrasts and LLM-generated synthetic data highlights the importance of inductive bias in contrastive learning. Standardized clinical questionnaires, such as the BDI, are the product of decades of psychometric validation, with response options crafted to be precise, mutually informative, and clinically interpretable. These properties appear to provide a stronger learning signal than synthetic sentences generated by LLMs, which may introduce unintended ambiguity or overlap despite explicit prompting. More broadly, this result points to a promising research direction: leveraging expert-designed instruments in other domains rather than mental health with structured assessment tools.

9 Conclusion and Future Work

Automatic depression detection has evolved from binary classification to fine-grained symptom identification, which is more useful for clinical practice. Nevertheless, even state-of-the-art retrieval models struggle to accurately distinguish between relevant sentences corresponding to topically similar symptoms. In this paper, we have introduced `BDI-batch`, a BDI-based contrastive learning approach that effectively selects contrastive pairs to enhance sentence-level representations. Our study introduced a novel way of leveraging standardized clinical questionnaires for computational mental health research. As future work, we plan to investigate the performance of our model in cross-disorder settings, e.g. evaluating whether a model adjusted for depression can correctly identify fine-grained symptoms in other disorders.

Limitations

Our experiments rely on publicly available datasets collected from a social media platform whose user base might not be demographically representative. This may limit the extent to which our findings generalize to broader populations, including older adults or communities that are less active online. Moreover, the study focuses exclusively on English-language content, which restricts conclusions to the linguistic and cultural context in which the data were produced.

Despite these constraints, our approach is highly adaptable and could be extended to other languages, platforms, and disorders. This flexibility opens opportunities to examine how symptoms manifest in different cultural and social settings. In future work, we plan to broaden the scope of our evaluation by incorporating multilingual data, alternative social media sources, and additional mental health conditions.

A further limitation of our approach is its reliance on the Beck Depression Inventory as the sole source of clinically grounded symptom definitions. While the BDI is a widely validated and influential instrument, its specific phrasing and conceptualization of symptoms inevitably shape the contrastive signal used during training, and the resulting representations may not transfer seamlessly to alternative questionnaires or diagnostic frameworks that operationalize depressive symptoms differently. Broader validation across other questionnaires, disorders, and symptom-grouping criteria remains an important direction for future work. The proposed framework could be applied to other clinically validated questionnaires with structured response options. Although our empirical evaluation is constrained by the limited availability of collections with sentence-level symptom relevance annotations, the method itself is questionnaire-agnostic and can be readily extended to additional datasets as they become available.

Moreover, although our method improves sentence-level retrieval performance, these gains should not be interpreted as evidence of clinical validity or diagnostic utility. The proposed models are intended to support information retrieval and research-oriented analyses, rather than to serve as tools for screening, diagnosis, or clinical decision-making without expert oversight.

Ethical considerations

This work aims to contribute to methods that support the understanding of mental health conditions, rather than serving as a diagnostic tool for targeting individuals. As noted by previous research, such computational approaches should be understood as tools that complement, rather than replace, clinical expertise (Neuman et al., 2012). Automated screening systems are best viewed as digital support mechanisms that may help reduce the burden on healthcare services. This study did not involve any interaction with social media users, nor did it provide any form of health advice. All analyses were conducted on publicly available data, in accordance with the platform's access policies and crawling guidelines. Because the data were publicly accessible and no user contact occurred, the study qualified for exemption from institutional ethics review.

Despite the use of publicly available data, analyzing online text for mental health–related signals raises broader ethical concerns regarding contextual integrity, as users may not anticipate that their posts could be repurposed for psychological analysis. Moreover, techniques developed for symptom retrieval could be misused in settings such as surveillance or profiling, which falls outside the intent of this work. Any application of such methods in real-world contexts would therefore require careful governance, transparency, and human-in-the-loop oversight, as well as adherence to relevant ethical, legal, and clinical standards.

Acknowledgments

This work was supported by MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22 & PID2025-167749OB-C21, co-funded by ERDF), and HUMANAIZE (AIA2025-163322-C62); and by the Xunta de Galicia – Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04 and ED431C 2026/52, co-funded by ERDF). We also acknowledge the project "Cátedra de IA aplicada a la Medicina Personalizada de Precisión" (Cátedras ENIA, TSI-100932-2023-3), funded by the Ministerio de Transformación Digital y Función Pública (Secretaría de Estado de Digitalización e Inteligencia Artificial) and NextGenerationEU.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Beng Heng Ang, Sujatha Das Gollapalli, and See-Kiong Ng. 2024. Nus-ids@ erisk2024: ranking sentences for depression symptoms using early maladaptive schemas and ensembles. *Working Notes of CLEF*, pages 9–12.

American Psychiatric Association APA. 2022. Diagnostic and statistical manual of mental disorders: DSM-5. *American psychiatric association*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

AP Bascuñana and Isabel Segura Bedmar. 2024. Apbuc3m at erisk 2024: natural language processing and deep learning for the early detection of mental disorders. *Working Notes of CLEF*, pages 9–12.

Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. 1996a. Comparison of Beck depression inventories-IA and-II in psychiatric outpatients. *Journal of personality assessment*, 67(3):588–597.

Aaron T Beck, Robert A Steer, and Gregory K Brown. 1996b. Beck depression inventory manual. the psychological corporation. *San Antonio, TX*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Ana-Maria Bucur. 2023. Utilizing chatgpt generated data to retrieve depression symptoms from social media. *arXiv preprint arXiv:2307.02313*.

Ana-Maria Bucur, Adrian Cosma, and Liviu P Dinu. 2021. Early risk detection of pathological gambling, self-harm and depression using bert. *arXiv preprint arXiv:2106.16175*.

Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P Dinu. 2023. It's just a matter of time: Detecting depression with time-enriched multimodal transformers. In *European conference on information retrieval*, pages 200–215. Springer.

Fabio Crestani, David E Losada, and Javier Parapar. 2022. Early detection of mental health disorders by social media monitoring. *Studies in Computational Intelligence*, 1018:4.

Sara Evans-Lacko, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Corina Benjet, Ronny Bruffaerts, Wai-Tat Chiu, Silvia Florescu, Giovanni de Girolamo, Oye Gureje, et al. 2018. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO world mental health (WMH) surveys. *Psychological medicine*, 48(9):1560–1571.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Johanna Habicht, Sruthi Viswanathan, Ben Carrington, Tobias U Hauser, Ross Harper, and Max Rollwage. 2024. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nature medicine*, 30(2):595–602.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. *spaCy: Industrial-strength natural language processing in Python*. Zenodo, Honolulu, HI, USA.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Ruijie Jiang, Thuan Nguyen, Prakash Ishwar, and Shuchin Aeron. 2024. Supervised contrastive learning with hard negative samples. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jaehoon Kim, Seungwan Jin, Sohyun Park, Someen Park, and Kyungsik Han. 2024. Label-aware hard negative sampling strategies with momentum contrastive learning for implicit hate speech detection. *arXiv preprint arXiv:2406.07886*.

Wenxiao Liu, Zihong Yang, Chaozhuo Li, Zijin Hong, Jianfeng Ma, Zhiquan Liu, Litian Zhang, and Feiran Huang. 2024. HNCSE: Advancing sentence embeddings via hybrid contrastive learning with hard negatives. *arXiv preprint arXiv:2411.12156*.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. erisk 2020: Self-harm and depression challenges. In *European conference on information retrieval*, pages 557–563. Springer.

Rodrigo Martínez-Castaño, Amal Htait, Leif Azzopardi, and Yashar Moshfeghi. 2020. Early risk detection of self-harm and depression severity using BERT-based transformers: ilab at CLEF eRisk 2020.

Alexa Mazur, Harrison Costantino, Prentice Tom, Michael P Wilson, and Ronald G Thompson. 2025. Evaluation of an ai-based voice biomarker tool to detect signals consistent with moderate to severe depression. *The Annals of Family Medicine*, 23(1):60–65.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Yair Neuman, Yohai Cohen, Dan Assaf, and Gabbi Kedma. 2012. Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine*, 56(1):19–25.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. Overview of erisk at CLEF 2021: Early risk prediction on the internet (extended overview). *CLEF (Working Notes)*, 1:864–887.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. erisk 2023: Depression, pathological gambling, and eating disorder challenges. In *European Conference on Information Retrieval*, pages 585–592. Springer.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2024. erisk 2024: Depression, anorexia, and eating disorder challenges. In *European Conference on Information Retrieval*, pages 474–481. Springer.

Javier Parapar, Anxo Perez, Xi Wang, and Fabio Crestani. 2025. Overview of erisk 2025: Early risk prediction on the internet. In *International conference of the cross-language evaluation forum for European languages*, pages 242–265. Springer.

Anxo Pérez, Marcos Fernández-Pichel, Javier Parapar, and David E Losada. 2025. Depresym: A depression symptom annotated corpus and the role of large language models as assessors of psychological markers. *Language Resources and Evaluation*, pages 1–26.

Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar, and Iryna Gurevych. 2023. Semantic similarity models for depression severity estimation. *arXiv preprint arXiv:2211.07624 To appear in: EMNLP 2023*.

Rafał Poświata and Michał Perełkiewicz. 2022. Opi@ lt-edi-acl2022: Detecting signs of depression from social media text using roberta pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282.

Anxo Pérez, Javier Parapar, and Álvaro Barreiro. 2022. Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine*, 132:102380.

Guozheng Rao, Yue Zhang, Li Zhang, Qing Cong, and Zhiyong Feng. 2020. MGL-CNN: a hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, 8:32395–32403.

Esteban A Ríssola, David E Losada, and Fabio Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, 2(2):1–31.

Ruba Skaik and Diana Inkpen. 2020. Using twitter social media for depression detection in the canadian population. In *Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference*, pages 109–114.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, NeurIPS 2021*, Online Event.

Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126.

World Health Organization WHO. 2025. Depressive disorder (depression). [Online; accessed 21-October-2025].

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022a. Symptom identification for interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9970–9985, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. Zhu. 2022b. Psychiatric scale guided risky post screening for early detection of depression. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 5220–5226, Vienna, Austria.

A First-Person Heuristic Impact

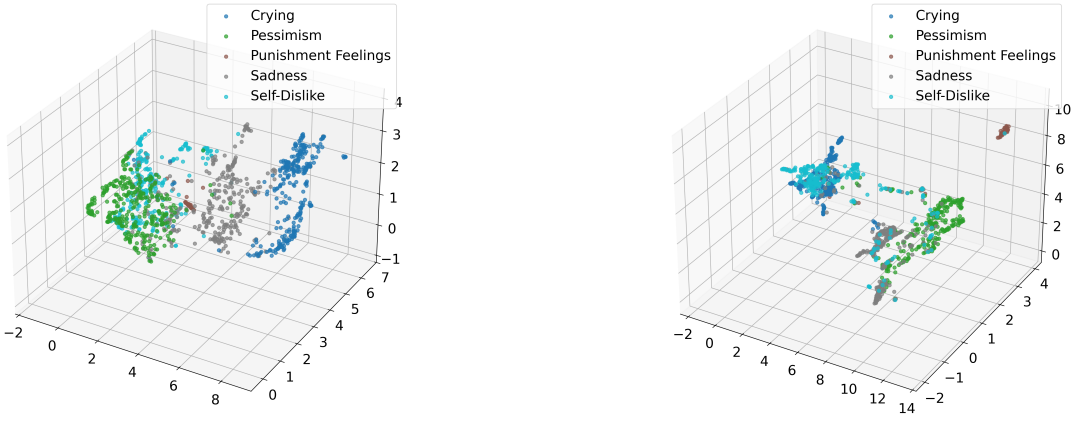
As shown in Table 7, the first-person promotion post-processing step improves performance for both the base and fine-tuned bi-encoder models. Nevertheless, contrastive learning fine-tuning provides further performance gains beyond those achieved with this heuristic alone.

B Embedding visualizations

We reduced the embedding dimensionality using UMAP (McInnes et al., 2018) with three output dimensions, 15 neighbors, and a minimum distance of 0.1. Figure 3 plots five related BDI symptoms and how the base model struggles to cluster them appropriately, with particularly poor separation between "*Self-dislike*" and "*Pessimism*". Our fine-tuned model yields a clearer separation. For instance "*Punishment Feelings*" relevant sentences form a coherent cluster. Still, the fine-tuned model fails to fully separate "*Self-dislike*" from "*Crying*". This highlights the intrinsic difficulty of distinguishing closely related symptoms. To view the full set of BDI symptoms, the reader can examine Figure 1 (the first 10 BDI symptoms) and Figure 4 (remaining 11 symptoms).

Model	R@100	NDCG@10	NDCG@1000
all-mpnet-base-v2 (w/o 1st person)	.261	.708	.674
all-mpnet-base-v2 (with 1st person)	.291	.834	.698
bdi-all-mpnet-base-v2 (w/o 1st person)	.327	.895	.797
bdi-all-mpnet-base-v2 (with 1st person)	**.372**	**.947**	**.817**

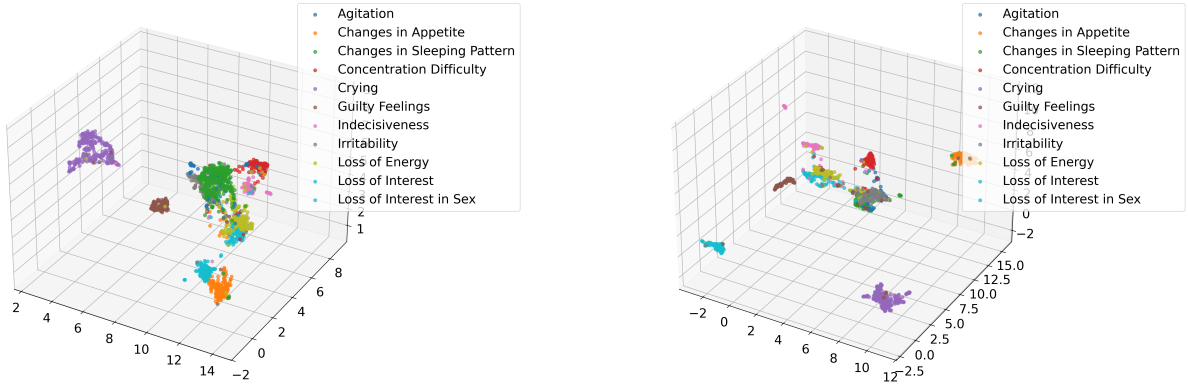
Table 7: Effect of the first-person promotion heuristic on the bi-encoders' retrieval performance.



(a) Base

(b) BDI fine-tuned

Figure 3: Sentences annotated as relevant to five related BDI symptoms. The left-side graph plots the UMAP representations of their `all-mpnet-base-v2` embeddings while the right-side graph plots the UMAP representations of our BDI-based variant.



(a) Base

(b) BDI fine-tuned

Figure 4: Sentences annotated as relevant to the remaining eleven BDI symptoms. The left-side graph plots the UMAP representations of their `all-mpnet-base-v2` embeddings while the right-side graph plots the UMAP representations of our BDI-based variant.

C Prompt for generating synthetic examples

Figure 5 shows the prompt utilized to generate synthetic examples.

For the 21 symptoms of the Beck Depression Inventory (BDI) questionnaire, I want you to generate 4 examples of sentences for each symptom. The sentences need to illustrate a person experiencing the symptom and talking about themselves, they can be either positive (having the symptom) or negative (not having the symptom).
For example, for Pessimism symptom sentences could be "I feel hopeless all the time" or "I am not pessimistic at all".' Just answer with a JSON in which the key is the symptom name and the values a list of four sentences. There cannot be overlapping between symptoms. Sentences should illustrate only one condition.

Figure 5: Prompt for generating synthetic examples.