Language Resources and Evaluation (2025) 59:2737–2762
https://doi.org/10.1007/s10579-025-09831-6

ORIGINAL PAPER

DepreSym: A Depression Symptom Annotated Corpus and the Role of Large Language Models as Assessors of Psychological Markers

Anxo Pérez[1] · Marcos Fernández-Pichel[2] · Javier Parapar[1] · David E. Losada[2]

Accepted: 16 April 2025 / Published online: 3 May 2025

Abstract

Computational methods for depression detection aim to mine traces of depression from online publications posted by Internet users. However, solutions trained on existing collections exhibit limited generalisation and interpretability. To tackle these issues, recent studies have shown that identifying specific depressive symptoms can lead to more robust and effective models. The eRisk initiative fosters research on this area and has recently proposed a new ranking task focused on developing search methods to find sentences related to depressive symptoms. This search challenge relies on the symptoms specified by the Beck Depression Inventory-II (BDI-II), a questionnaire widely used in clinical practice. It includes symptoms such as sadness, irritability or lack of sleep. Given the input submitted by systems participating in eRisk, we first apply top-k pooling over the systems' relevance rankings, obtaining a diverse set of sentences. These sentences are judged for relevance, leading to *DepreSym*, a dataset consisting of 21,580 sentences annotated according to their relevance to the 21 BDI-II symptoms. This dataset serves as a valuable resource for advancing the development of models that monitor depression markers. Due to the complex nature of this relevance annotation, we designed a robust assessment methodology carried out by three expert assessors, including a trained psychologist. As part of this study, we explore the potential of recent Large Language Models (ChatGPT, GPT4 and Vicuna) as assessors in this complex task. We undertake a comprehensive examination of the LLMs' performance, studying their main limitations and analysing their role as a complement or replacement for human annotators. Finally, we incorporate our dataset into the Benchmarking Information Retrieval (BEIR) framework for a thorough search evaluation. We use state-of-the-art retrieval systems, including lexical, sparse, dense and re-ranking architectures, to gain insights about the dataset's complexity and identify potential avenues for improvement.

Keywords Depression · Social media mining · Large language models · Search · Information retrieval

Extended author information available on the last page of the article

1 Introduction

Inspired by clinical practice, there has been a growing interest in designing automated predictive models that focus on identifying specific depressive symptoms. These approaches diverge from traditional depression screening models that rely on the presence of general markers (e.g., word counts, emotion levels, posting hours), which offer less personalised and interpretable solutions (Harrigian et al., 2020; Walsh et al., 2020). Recent studies have shown the potential of symptom-based detection models (Nguyen et al., 2022b; Pérez et al., 2022, 2023; Zhang et al., 2022a, c).

In this context, the CLEF eRisk depression severity task (Losada et al., 2019) released the first collection of data that contains symptom-level evidence that was obtained directly from individuals. The goal of the task was to estimate the severity of the 21 symptoms enumerated in the BDI-II (Beck et al., 1996). The BDI-II is a standardised questionnaire covering symptoms such as pessimism, suicidal ideas or sleep problems. Each symptom is handled by a BDI-II question, which has four possible closed-text answers. The answers are organised in an ordinal scale that represents the severity of the symptom. The eRisk's organisers designed a *human-in-the-loop* approach, where real responses to the BDI-II were provided by Reddit users. These users also gave consent to access their history of publications on the platform. This innovative challenge aimed to foster the design of automatic systems that analyse the thread of users' messages and predict the user's responses to the BDI-II (Losada et al., 2019). However, the ground truth (users' responses to the BDI-II) was at user level. Thus, no explicit association existed between BDI-II symptoms and specific textual extracts from the publications. Two recent studies have attempted to fill this void by constructing fine-grained datasets that label depressive symptoms at sentence level (Pérez et al., 2023; Zhang et al., 2022a). This paper contributes to this line of research by introducing *DepreSym*, a dataset to encourage the development of models that rely on symptom-level screening of depression. *DepreSym* consists of 21,580 sentences that are labelled in terms of their relevance to the BDI-II symptoms.[1] This represents the largest symptom-level dataset to date. To construct our dataset, three expert assessors annotated a pool of sentences associated with each symptom. The candidate sentences were obtained using top-k pooling from relevance rankings produced by participants in the CLEF 2023 eRisk Lab.[2] A total number of 37 different ranking methods were considered. Pooling over these search methods helps to increase the diversity of the candidate sentences.

The assessors were instructed to consider a candidate sentence as relevant if (i) it is on-topic and, additionally, (ii) provides explicit information about the individual's condition in relation to the symptom. This two-side notion of relevance is more complex compared to standard relevance annotation, and it required us to develop a robust annotation methodology with detailed assessment guidelines. To validate the

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

[2] https://early.irlab.org/.

effectiveness of our methodology, we calculate the inter-rater agreement and conduct further analysis of the resulting set of judgements. Additionally, we explore here the ability of recent state-of-the-art conversational AI systems based on Large Language Models (LLMs) to annotate the dataset. Specifically, we employ the latest versions of ChatGPT (Forbes, 2022), GPT-4 (OpenAI, 2023) and Vicuna (Chiang et al., 2023) as complex relevance assessors. One of the main advantages of LLMs is their ability to accurately process large amounts of data and, thus, they can significantly reduce the time and effort required for manual assessment. Comparing the performance of LLMs with human assessors provides insights into the strengths and limitations of both approaches. Human assessors are considered the gold standard for relevance assessment, but are also subject to biases and errors that can affect their performance. Examining the performance of LLMs in relation to humans can help us to understand how well these models can replicate human behaviour. Finally, in an effort to thoroughly evaluate and better understand the difficulty of our dataset, we integrate it into the well-known Benchmarking-IR (BEIR) framework (Thakur et al., 2021). This integration allows us to leverage a wide range of state-of-the-art retrieval systems, spanning from lexical and sparse to dense and re-ranking retrieval architectures. This empirical validation helps to provide a complete view of the dataset's main search challenges. The incorporation of our dataset into BEIR enriches this test bank. As a matter of fact, BEIR covers a wide range of tasks and domains but, still, there is a need to improve up-on the benchmark by including test collections with other length granularities and task-oriented notions of relevance. In this respect, DepreSym represents a good addition to the existing set of BEIR collections. Our main contributions are:

- *DepreSym*, a dataset of 21,580 sentences annotated for relevance with respect to the 21 BDI-II depression symptoms. This resource, developed with top-k pooling and expert annotation, stands out for its quality and potential in mental health research.
- An examination of the strengths and limitations of the latest conversational LLMs for annotating standardised depression symptoms. Our study also explores how the LLMs' assessments influence the rankings of retrieval systems.
- A thorough evaluation of *DepreSym* using state-of-the-art retrieval systems within the BEIR framework (Thakur et al., 2021). This includes a focused symptom-by-symptom analysis, identifying performance variances and avenues for future improvements.

2 Related work

2.1 Depression Symptoms and Social Media

Recent studies on social media analysis have leveraged the extraction of fine-grained depressive symptoms to improve mental health models (Coppersmith et al., 2018; Nguyen et al., 2022b; Pérez et al., 2022, 2023; Zhang et al., 2022c). Such symptom-level extraction has demonstrated potential to improve performance, generalisation and interpretability. The development of these methods builds upon multiple previous studies that analysed the interplay between language and mental states (Chancellor and De Choudhury, 2020; Crestani et al., 2022; Losada and Gamallo, 2020; Pennebaker, 2011; Ríssola et al., 2021). This class of social media monitoring studies have led to the production of corpora that help in preventing mental health disorders (Santos et al., 2023) or identifying related aspects, such as emotions or certain types of speech (Babakov et al., 2023).

A significant initiative in the area of digital mental health is the *Early Risk Prediction On The Internet* lab (eRisk[3]). Since 2017, the eRisk initiative has been running under the Conference and Labs of the Evaluation Forum (CLEF) evaluation campaign. The organisers of eRisk have proposed different shared-data challenges related to specific aspects of social media risk assessment. In eRisk 2019, 2020 and 2021, a new task on *Measuring the Severity of the Signs of Depression* was proposed. The task aims to estimate the level of depression of a given social media user from its Reddit[4] publications. The target levels correspond to the BDI-II questionnaire scores (Beck et al., 1996). The BDI-II[5] is a self-report instrument designed to assess the severity of depressive symptoms in adolescents and adults. It consists of 21 symptoms that measure attitudes and symptoms of clinical depression, such as agitation, sadness or loss of energy (Steer et al., 1986). The participating systems in eRisk 2019–2021 were given the posting history of multiple test users and they were asked to predict the user's responses to the BDI-II questionnaire. The predicted answers had to be based on the evidence found in the history of publications. The specifics of all contributing systems can be consulted in the eRisk overviews (Losada et al., 2019, 2020; Parapar et al., 2021).

As this eRisk task gained traction, the exploitation of symptoms for developing depression detection models received increasing attention. Instead of relying on vast amounts of unstructured data to perform two-class categorisation (depression vs non-depression), some researchers started to see the potential of structured clinical questionnaires and symptom's lists. A recent line of work has focused on developing models that integrate depressive symptoms as reliable clinical markers. These models are driven by the need of discovering reliable depression markers that help health professionals in their diagnosis (Coppersmith et al., 2018).

Many of these studies employed LLMs as the core technology for classification (Nguyen et al., 2022a; Pérez et al., 2022, 2023; Zhang et al., 2022b, c). For instance, Zhang et al. (2022c) introduced a psychiatric scale-guided method to screen risky posts (related to dimensions defined in clinical depression questionnaires). By using templates from BDI-II, these authors obtained direct expressions of depressive symptoms. A hierarchical network incorporating BERT further aggregated the selected user posts and assigned higher weights to key contents related to depressive symptoms. The authors used the eRisk2017 dataset for evaluation, comparing their approach against several baselines. This study

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

[4] https://www.reddit.com/.

[5] The full BDI-II can be consulted at https://erisk.irlab.org/2019/ (Task 3).

underscored the potential of integrating the screening of posts exhibiting depressive symptoms with the use of LLMs, thereby facilitating the emission of accurate predictions. Nguyen et al. (2022a) also investigated BERT-based methods for detecting depressive symptoms using nine indicators from the PHQ-9 scale. This team used heuristics and regular expression patterns to construct weakly-supervised data from Reddit. Their strategy involved two models: one targeting PHQ-9 symptoms using manually crafted patterns and another broader depression detection model based on the frequency of the patterns in user posts. Under in-domain experiments, these constrained models performed competitively compared to a standard unconstrained BERT classifier. The performance of this approach was also evaluated under a *transfer learning* setting, covering three different datasets (RSDD (MacAvaney et al., 2018), eRisk2018 (Losada et al., 2018) and the TRT corpus (Wolohan et al., 2018)). The classifiers performed well compared to a standard depression classifier and generalised better to other datasets. Moreover, the models highlighted social media extracts related to depressive symptoms, a clear demonstration of their interpretability.

A predominant limitation in the existing literature on symptom detection is the lack of high-quality training data. The absence of large-scale annotated corpora made that previous research efforts had to rely on unsupervised or weakly supervised methods, guided by pattern matching. A notable exception is the work done by Zhang et al. (2022b) on symptom-based models for mental health detection. This team introduced PsySym, the first symptom-based dataset with manual annotations of 38 symptoms related to 7 mental disorders. The authors established symptom classes according to the DSM-V (Nuckols and Nuckols, 2013), and they leveraged symptom descriptions on different clinical questionnaires as a resource to obtain candidate posts to annotate. First, the authors searched in Reddit for posts that are deemed to express symptoms related to mental disorders. Candidate posts were only selected from mental health-related subreddits, where most posts are likely to be relevant. Their matching approach leveraged embedding-based retrieval methods (Reimers and Gurevych, 2019) (instead of keyword matching) to get the candidate sentences for annotation. Next, the annotation process was performed via crowdsourcing. Given the labelled data, the authors proposed a methodology for mental health detection using supervised learning models. After training symptom-level classifiers, the resulting prediction probabilities served as the base for building a vectorial representation of symptoms. This led to 38-dimensional feature vectors that outperformed all pure-text methods, including a solid BERT model. These results showed the potential of symptom-based methods to help in precision diagnosis and mental health screening.

Following PsySym, another study presented the BDI-Sen (Pérez et al., 2023), the first symptom-based dataset that covers all the symptoms from the BDI-II. It contains a total of 4973 annotated sentences. Similar to PsySym, the authors identified sentences that were estimated to be relevant to BDI-II symptoms. This candidate selection was supported by embedding-based retrieval. Next, three expert assessors manually annotated the candidates. Only 17% of candidates were labelled as positive, underscoring the difficulty in extracting sentences linked to depressive symptoms. Furthermore, an in-depth symptom-by-symptom analysis revealed significant linguistic and emotional differences between positive and negative sentences. The authors also conducted experiments for two different classification tasks, namely: symptom detection and symptom severity classification. These experiments were supported by different LLMs. In the symptom detection task, the trained models proved to be effective at finding relevant sentences. Moreover, in subsequent experiments, these models demonstrated robust generalisation abilities in detecting symptoms related to other diseases.

2.2 LLMs as relevance annotators

High-quality assessments are essential to obtain accurate and reliable benchmarks (Büttcher et al., 2007). Annotations need to be consistent, unbiased, and representative of the task at hand. Low-quality assessments potentially lead to inaccurate evaluations and unreliable conclusions (Scholer et al., 2011). The process of manually annotating test collections requires significant human effort, frequently requiring domain experts. As a consequence, a number of steps have been taken to reduce the cost and biases of the labelling process (Moghadasi et al., 2013; Sakai, 2009).

With the incredible development of LLMs, a potential application of these models is to assist in tasks such as relevance labelling. This represents a natural application of these intelligent agents, as was the replacement of TREC annotators by crowdsourcing (Alonso and Mizzaro, 2009). Initial steps towards this goal were taken in Gilardi et al. (2023), a study that demonstrated that ChatGPT outperforms crowd-workers for a tweet annotation task. Other researchers focused their efforts on improving annotation through prompt engineering (He et al., 2023). An evaluation of the accuracy of LLMs for annotating two TREC test collections was presented in Faggioli et al. (2023). Previous efforts (Meyer et al., 2022) were oriented to produce synthetic training data for a conversational agent in the context of behaviour change. In our study, we intend to go one step further by evaluating the most recent LLMs for a highly demanding annotation task. Specifically, we put these conversational AI agents under scrutiny for assessing the relevance of sentences in relation to specific BDI-II symptoms. Thus, we are considering a scenario where the notion of relevance is complex. A relevant sentence has to be on-topic but, additionally, it should convey evidence about the individual's condition with respect to the BDI symptom. Moreover, the context is short (judgements at sentence-level) and, to study this effect, we analyse the agreement between human annotations, including those coming from experts in the field, and machine-generated annotations.

2.3 Ranking methods

Information Retrieval (IR), a pivotal branch of Computer Science, focuses on retrieving relevant contents in response to a specific query or information need. In our study, we derive queries from the BDI-II questionnaire and consider the sentences in the DepreSym dataset as candidate retrieval units. Traditional lexical retrieval approaches, such as BM25, are based on exact term matching between the query and the candidate texts (Robertson et al., 1995). However, these approaches exhibit limitations in capturing semantic similarities between terms (Berger et al., 2000). The last decade has witnessed the appearance of new retrieval models that leverage deep NLP models and consider contextual information (Devlin et al., 2018; Vaswani et al., 2017). Among these new retrieval strategies, we can find sparse models, which still maintain some sort of traditional matching. For example, docT5query (Nogueira et al., 2019) augments documents using a sequence-to-sequence model for subsequent lexical retrieval. DeepCT (Dai and Callan, 2020) employs BERT to generate dense document representations prior to retrieval. Dense retrieval approaches like ANCE (Xiong et al., 2020) employ vector space mappings of queries and documents to facilitate retrieval based on similarity metrics. In our research, we exploit BEIR's toolkit of models (Thakur et al., 2021) for conducting retrieval experiments against DepreSym. These search tests aim to shed light on the difficulty of finding traces of depression symptoms.

3 DepreSym

This section describes the construction of *DepreSym*, a resource that builds on Task 1 from the eRisk 2023 Lab (Parapar et al., 2023). This is a novel task that consists of identifying sentences that are indicative of the presence of clinical symptoms in the individuals who posted these sentences. We follow the BDI-II, a well-studied clinical questionnaire covering 21 symptoms of depression. It includes emotional (e.g., *Pessimism or Sadness*), cognitive (e.g., *Indecision*) and physical (e.g., *Fatigue*) symptoms (Beck et al., 1996). Table 1 shows four BDI-II symptoms and their possible responses. The eRisk dataset consists of sentences written by multiple social media users (from the Reddit platform). The original user posts were segmented into sentences and a TREC-style collection was created (3,807,115 sentences from 3107 unique users). All extracted sentences were public and Reddit terms allow the use of its contents for research purposes.

Table 1 Four BDI-II symptoms (*sadness*, *low energy*, *self-dislike* and *guiltiness*) and their associated options with their descriptions

Sadness	Low energy
0. I do not feel sad	0. I have as much energy as ever
1. I feel sad much of the time	1. I have less energy than I used to have
2. I am sad all the time	2. I do not have enough energy to do very much
3. I am so sad or unhappy that I can't stand it	3. I do not have enough energy to do anything
Self-dislike	Guiltiness
0. I don't feel disappointed in myself	0. I don't feel particularly guilty
1. I am dissapointed in myself	1. I feel guilty a good part of the time
2. I am disguted with myselfh	2. I feel quite guilty most of the time
3. I hate myself	3. I feel guilty all of the time

Table 2 Examples of sentences for the symptom *Loss of Energy*. Sentences are paraphrased for anonymity purposes

Relevance	Sentence
0	"*Learn new ideas consumes energy, but builds neural connections*"
	"*Low electrolytes can cause a person to feel low on energy*"
1	"*Even brushing my teeth is too exhausting for me right now*"
	"*I became constantly lethargic, drowsy, and unable to concentrate*"

The eRisk participants were given the entire collection of sentences and were asked to submit 21 rankings of sentences (one for each BDI-II symptom) ordered by decreasing relevance to the symptom. Each participating team could submit up to 5 variants (runs) and each ranking had up to 1000 sentences. Prior to annotation, we obtained candidate sentences by following a top-k ($k = 50$) pooling approach on the submitted runs (37 runs from 10 different teams). Table 2 provides two examples of candidate sentences annotated as non-relevant (0) and relevant (1) for the symptom *Loss of energy*. Note that all sentences are somehow on-topic but only those in the lower block were labelled as relevant. These two relevant sentences offer insights into the individual's state related to the BDI-II symptom. This stringent notion of relevance adds complexity to the labelling process.

For each symptom, the upper block of Table 3 reports the total number of annotated sentences (first row), and the number of sentences marked as relevant (second row). The pool sizes range from 829 to 1150 sentences, with an average of 1028 per symptom. Here, relevant sentences are those unanimously agreed upon by all human assessors. The number of relevant sentences is low, with a mean of about 11% relevant sentences in the pool of candidates. The number of relevant sentences ranges from 21 to 260, averaging 128 per symptom. The rest of the blocks report the annotation agreement over the set of symptoms. Further details about the assessment process will be presented next.

4 Human annotation

To ensure a consistent evaluation, we designed a set of instructions that guide the assessment process.[6] The guidelines were given to the human annotators and, additionally, these textual instructions were used to prompt the LLMs in our study of automatic judgements. A sentence should be considered relevant only if it provides "information about the individual's state related to the BDI-II symptom".

We selected three human assessors with different backgrounds: a field expert (with background in Psychology), a PhD student and a Postdoc (both of them with background in Computer Science). This diversity allows us to investigate the assessments done by a psychologist with clinical experience and compare them with those

[6] https://erisk.irlab.org/guidelines_erisk23_task1.html.

Table 3 Number of sentences and annotation agreement (in percentage)

	Sadness	Pessimism	Sense of failure	Loss of pleasure	Guiltiness	Self-Punishment	Self-dislike	Self-incrimination	Suicidal ideas	Crying	Agitation
# Sentences	1110	1150	973	1013	829	1079	1005	1072	953	983	1080
# Rel. sent.	179	104	160	97	83	21	158	76	260	230	69
GPT-4 vs Cons.	72.6	75.7	73.6	74.6	75.6	88.2	72.0	75.6	77.2	79.8	80.7
GPT-4 vs Maj.	76.1	77.1	76.0	82.0	81.2	89.3	81.6	81.3	85.9	87.7	80.7
PhD s. vs Rest	80.0	70.5	84.2	88.5	90.8	97.0	81.4	86.5	87.4	89.3	88.9
GPT-4 vs Rest	73.0	75.6	75.3	79.6	78.9	89.6	73.3	76.6	77.4	82.8	81.4
Psy. vs Rest	83.0	76.8	74.1	82.7	90.0	95.8	85.3	86.7	89.8	88.1	88.4
GPT-4 vs Rest	73.6	75.6	73.9	74.5	76.2	88.2	74.2	77.4	81.7	82.0	80.0
Postdoc vs Rest	85.4	78.3	84.5	82.1	86.9	92.4	84.2	86.2	89.4	88.9	87.6
GPT-4 vs Rest	74.8	77.3	74.0	77.2	77.3	88.4	78.1	78.5	81.2	82.5	80.7
	Social issues	Indecision	Worthlesness	Low energy	Sleep issues	Irritability	Appetite issues	Concentration	Fatigue	Low libido	Avg.
# Sentences	1077	1110	1067	1082	938	1047	984	1024	1033	971	1028
# Rel. sent.	70	61	71	129	203	94	103	83	123	97	118
GPT-4 vs Cons.	75.9	79.2	80.9	78.3	62.1	84.1	71.7	83.9	74.8	81.6	–
GPT-4 vs Maj.	81.2	83.3	84.7	84.5	76.4	88.0	80.5	88.2	83.7	85.6	

Table 3 (continued)

	Social issues	Indecision	Worthlesness	Low energy	Sleep issues	Irritability	Appetite issues	Concentration	Fatigue	Low libido	Avg.
PhD s. vs Rest	87.7	90.8	90.3	88.1	80.7	86.1	84.6	90.9	88.2	92.4	86.9
GPT-4 vs Rest	78.4	80.9	81.4	79.5	65.1	85.0	75.3	84.4	77.2	82.2	78.7
Psych. vs Rest	85.7	90.6	91.7	91.9	83.1	93.6	86.8	94.0	89.2	93.3	87.6
GPT-4 vs Rest	76.6	80.5	82.9	80.8	70.0	85.6	75.2	86.8	80.0	84.8	79.1
Postdoc vs Rest	86.4	88.3	89.6	90.5	79.2	92.5	80.9	91.7	86.5	86.0	86.5
GPT-4 vs Rest	78.0	80.3	82.3	80.8	65.4	85.5	73.3	84.8	76.3	81.8	79.0

Table 4 Agreement between each LLM and two types of ground truth obtained from human annotators (consensus and majority)

LLM	Prediction	Consensus			Majority		
		Rel	Not Rel	κ	Rel	Not Rel	κ
ChatGPT	Rel	**2358**	9277	0.18	**4241**	7394	0.38
	Not Rel	113	**9832**		290	**9655**	
GPT-4	Rel	**2296**	4755	0.38	**3916**	3135	0.57
	Not Rel	175	**14,354**		615	**13,914**	
Vicuna13B	Rel	**2159**	10,393	0.12	**3916**	8636	0.22
	Not Rel	312	**8716**		615	**8413**	

Bold values represent correct predictions

made by humans with more limited domain-specific knowledge. Initially, the assessors were tasked with labelling the sentence pools from the first three BDI-II topics. For this round, assessors were allowed to mark sentences as “undecided”. To estimate the agreement for these three symptoms, we calculated pairwise Cohen’s Kappa for single raters and Krippendorff’s Alpha for ordinal scales between all raters. The Kappa values ranged between 0.18 and 0.51 with a median value of 0.38, suggesting a low average agreement. The mean Krippendorff’s α was 0.32, which is below the generally accepted threshold ($\alpha \geq 0.667$) for reliable annotations (Krippendorff, 2018; McHugh, 2012).

After this initial assessment, we had a briefing with the three annotators to address any ambiguities and to promote more consistent assessments. Following this session, we requested them to re-evaluate the sentences from the initial three BDI-II symptoms. This time, the assessors could not assign “undecided” labels. The Cohen’s Kappa values ranged between 0.30 and 0.68, with a median value of 0.55. The average Krippendorff’s α improved, reaching 0.51, but still fell short of the desired threshold. After processing the entire set of 21 symptoms, the agreement analysis yielded Cohen’s Kappa values ranging between 0.58 and 0.65, with a median at 0.58, and a Krippendorff’s α of 0.60. These improved metrics, compared with those of the initial experiment, underscore the efficacy of the briefing with the assessors. However, the moderate agreement scores exemplify the inherent complexity of the task. As a result of this study, we generated two distinct types of relevance assessments: consensus and majority. We consider the consensus as a higher-quality set of annotations, capturing sentences with unambiguous relevance.

5 LLMs as automatic annotators

Given the strong capacities of current LLMs in multiple language understanding tasks, it is natural to wonder about their effectiveness for our stringent relevance assessment task. We prompted three LLMs (ChatGPT, GPT-4 and Vicuna13B) with the same instructions and example sentences provided to the human assessors and analyse here the agreement between human annotators and the LLMs (see Table 4). The analysis is conducted for the two classes of ground truth annotations: (i) *Consensus*, where relevant sentences were marked as so by all human assessors, and (ii) *Majority*, where relevant sentences are those marked as so by at least two human annotators.

Under the consensus ground truth, ChatGPT accurately identified 95% of the relevant sentences (2358 out of 2471), compared to 93% for GPT-4 (2296 out of 2471) and 87% for Vicuna (2159 out of 2741). However, all the models models struggled to correctly identify sentences marked as non-relevant (accuracies of 51, 75 and 46% for ChatGPT, GPT-4 and Vicuna, respectively). This trend persists for the majority ground truth, but, in this case, the correlation with human judgements shows a pronounced improvement. With majority, the Cohen's κ level of agreement increases from 0.18 to 0.38 for ChatGPT, from 0.38 to 0.57 for GPT-4 and from 0.12 to 0.22 for Vicuna. These figures indicate that GPT-4 was the best performing model, while Vicuna obtained the worst results. Another finding is that the "non-relevant" predictions of the models tend to be trustworthy. For instance, ChatGPT identified correctly 9832 out of 9945 non-relevant sentences. However, the predictions of relevance are much noisier, as we can see that the LLMs tend to label many non-relevant sentences as relevant. In the past, LLMs have been effective in categorising on-topic extracts. However, detecting sentences that inform about the psychological state of a given individual goes well beyond a standard relevance task and, according to our results, this imposes a barrier to current LLMs. These findings prompt us to discuss the potential of LLMs in a hybrid annotation approach, a topic we delve into in Sect. 5.4.

5.1 Symptom-based agreement

Table 3 displays the agreement statistics for each symptom. In the second block, we report the percentage of agreement between GPT-4 and the two classes of ground truth (Consensus and Majority). The agreement percentages are generally stable across all symptoms. These percentages are higher for majority (82.63% mean overall) compared to consensus (77.04%). Moreover, ChatGPT (65 and 56%, respectively) and Vicuna13B (71 and 51%, respectively) achieved substantially lower agreement values.[7]

We further conducted a pairwise comparison between the annotations provided by GPT-4 and each human annotator, reported in the last three blocks of Table 3. In this approach, our reference ground truth was derived from the consensus annotations of the two remaining human experts. For example, to compare the PhD student vs GPT-4, the ground truth of relevant sentences was obtained from both the postdoc and the psychologist. Across all such comparisons, each human annotator consistently outperformed GPT-4 in terms of agreement percentages. The only exception was in one symptom (*Pessimism*), where GPT-4 against the PhD student

[7] We only include GPT-4's results in Table 3 as it was the best performing LLM.

Table 5 Correlations between the official ranking of systems (consensus qrels) and the ranking of systems obtained from the qrels built with a single annotator

	Annotators			
	GPT-4	Psychologist	Postdoc	PhD student
Kendall τ	0.86	**0.98**	0.95	0.94
τ_{ap}	0.81	**0.97**	0.91	0.88

Bold values represent the best agreement

achieved a higher percentage of agreement (75.57% vs 70.52%, respectively). On average, human annotators exhibited agreement scores surpassing 85%, whereas GPT-4's scores remained under 80%. Interestingly, the psychologist was the human who produced superior agreement scores by a narrow margin.

5.2 Inter-rater agreement

The inter-rater agreement, measured using Cohen's κ between ChatGPT and the human annotators, ranged from 0.29 to 0.32, with a median value of 0.31. For GPT-4, the Cohen's κ scores ranged from 0.52 to 0.54, with a median of 0.53. Additionally, Krippendorff's α for the combination of the three human annotators and ChatGPT was 0.40, while Krippendorff's α with GPT-4 was 0.56. These results confirm the previous findings that GPT-4 is a more reliable annotator compared with ChatGPT.

5.3 Correlation of systems rankings

We also compared the official ranking of the 37 participating search systems in the eRisk task against a hypothetical ranking based on assessments from a single annotator. To that end, we ranked the systems by decreasing Mean Average Precision (MAP) and compared the rankings with Kendall's τ and AP Correlation (τ_{ap}[8]) Yilmaz et al. (2008). This analysis allows us to explore to what extent the use of a single annotator alters the system's rankings.

Looking at the results in Table 5, we can observe that GPT-4 yields a high correlation (0.86 and 0.81), although lower than the correlation levels achieved by the human annotators. Note that the human assessors were involved in the construction of the official qrels, while GPT-4 was not part of the official evaluation process. In this regard, the performance of GPT-4 might be underestimated. The results also suggest that the assessment effort could have been reduced by involving a single human assessor. Notably, the psychologist exhibits a nearly perfect correlation with the official consensus-based ranking (0.98). The correlations suggest a relative order among human annotators, Psych > Postdoc > PhD students, which is a natural consequence of their domain knowledge and level of experience.

[8] τ_{ap} assigns greater weight to errors made to the systems positioned higher in the ranking.

Looking at AP correlation (τ_{ap}), which penalises more the alterations at top-ranked positions, the trends remain consistent with those found with τ.

5.4 Final remarks

Our results indicate that LLMs demonstrate a significant superiority in identifying sentences deemed relevant according to the ground truth, as opposed to those deemed non-relevant. This finding deviates from the tendencies observed in prior research (Faggioli et al., 2023), wherein varying patterns emerged based on the specific dataset. We believe that our results give grounds to propose a new efficient hybrid labelling strategy, where LLMs act as filters that automatically remove non-relevant sentences from the pools. As shown in Table 4, the "non-relevance" predictions of LLMs are quite accurate and, thus, the human annotation effort could be reduced to review those sentences estimated as relevant by the LLM. Thus, GPT-4 would reduce the human workload by approximately 68%, eliminating the need to annotate around 15,000 sentences. Considering that the average human effort per assessor was 70 h (21,580 sentences), this reduction would save around 49 h of work per human. Furthermore, reducing the burden on human annotators could potentially lead to improved annotation quality and allow for an increase in the size of the annotation pool, allowing for more documents to be reviewed.

6 Search experiments

In this section, we evaluate different retrieval models applied to the DepreSym dataset. These experiments, which are supported by the Benchmarking-IR (BEIR) framework (Thakur et al., 2021), help to understand the difficulty of the BDI-based search challenge and, additionally, provide an assorted set of baseline variants that serve as a reference point for subsequent research in this field.

Our experimental design explores a wide range of state-of-the-art retrieval models, powered by diverse architectures, and covers the main classes of search solutions (lexical, sparse, dense, late interaction and re-ranking). The evaluation aims at investigating the inherent complexity of the DepreSym dataset. Furthermore, it allows us to study the effectiveness and challenges posed by general retrieval models when applied to this specialised context.

In order to search for sentences that are relevant to BDI-II symptoms, the formulation of queries is the essential initial step. Drawing inspiration from previous studies (Parapar et al., 2023; Pérez et al., 2023; Nguyen et al., 2022b; Zhang et al., 2022a), our query construction approach leverages the content of the BDI-II clinical questionnaire. Specifically, we exploit the textual descriptions linked to each BDI-II symptom (see Table 1). More specifically, we designed three distinct query construction strategies:

Symptom Title: This query type employs only the symptom's title (e.g., *Sadness* or *Loss of Energy*).

Symptom Title + Responses: Here, we concatenate the symptoms' title with its four possible responses, creating a long query.
Average of symptom responses: In this variant, we build individual queries, one for each response, resulting in four unique queries per symptom. In the tables, the performance score reported is the average effectiveness across these four queries.

In the following, we briefly discuss the retrieval models tested, according to their architecture:

- *Lexical Retriever*: Utilises a bag-of-words retrieval function through Anserini's BM25 implementation (Yang et al., 2017). It is based on BM25's default Lucene parameters.
- *Dense Retrievers*:
  - *Sentence Similarity-Based Models*. These strategies compute sentence similarity by comparing embeddings of queries and documents (sentences in our case).[9]
  - *Dense Passage Retrieval (DPR)*. This model (Karpukhin et al., 2020) adopts a bi-encoder architecture trained with hard negatives. It is specialised for vector-based retrieval and incorporates training from four QA datasets.
  - *ANCE*. Under a bi-encoder architecture, ANCE (Xiong et al., 2020) employs Approximate Nearest Neighbor (ANN) indexes to create robust training instances.
  - *TAS-B*. TAS-B (Hofstätter et al., 2021) combines Margin-MSE loss with in-batch negative loss functions for enhanced efficiency. It is trained with balanced Topic Aware Sampling.
- *Sparse Retrievers*:
  - *DeepCT*. This model (Dai and Callan, 2020) employs BERT-base-uncased models to learn term frequencies, creating pseudo-documents that augment keyword-based search, complemented by BM25 for the unmodified parts.
  - *SPARTA*. This search alternative (Zhao et al., 2020) estimates similarity scores using non-contextualised query embeddings from BERT and context-aware document embeddings.
  - *docT5query*. This document expansion technique (Nogueira et al., 2019) uses a T5-base model to generate synthetic queries appended to the original documents, followed by BM25-based retrieval.
- *Late Interaction:* ColBERT (Khattab and Zaharia, 2020) employs a late-interaction approach with a bag of contextualised token embeddings. It aggregates interactions through sum-max-pooling of query terms and dot-product across passage terms, trained on the MS MARCO dataset.

[9] We employed the top performing models on the semantic search SBERT's leaderboard: https://www.sbert.net/docs/pretrained_models.html.

Table 6 Search results for the DepreSym dataset

Query	Retriever	Model	AP			P			R			NDCG		
			@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
Symptom title	Dense	all-mpnet	0.005	0.054	$\underline{0.201}$	0.214	0.214	**0.278**	0.014	0.150	$\underline{0.680}$	0.183	0.270	0.532
		multi-qa-mpnet	0.005	0.058	0.197	0.233	$\underline{0.320}$	0.144	0.012	$\underline{0.165}$	0.667	0.212	$\underline{0.310}$	$\underline{0.534}$
		DPR	0.003	0.009	0.028	0.105	0.078	0.052	0.005	0.034	0.234	0.113	0.179	0.179
		ANCE	0.006	0.044	0.143	0.215	0.255	0.116	0.012	0.126	0.519	0.213	0.252	0.425
		TAS-b	$\underline{0.008}$	$\underline{0.059}$	0.169	0.257	0.296	0.123	0.014	0.148	0.575	0.251	0.293	0.469
	Lexical	Anserini BM25	0.006	0.038	0.110	0.190	0.200	0.086	0.011	0.107	0.400	0.200	0.204	0.333
	Sparse	DeepCT	0.005	0.037	0.108	0.200	0.200	0.085	0.012	0.106	0.395	0.203	0.201	0.328
		SPARTA	$\underline{0.008}$	0.051	0.160	$\underline{0.319}$	0.265	0.117	$\underline{0.018}$	0.136	0.523	$\underline{0.290}$	0.270	0.435
		docT5query	0.006	0.038	0.038	0.190	0.200	0.086	0.011	0.107	0.400	0.200	0.204	0.332
	Late interact.	ColBERT	0.005	0.037	0.130	0.176	0.233	0.112	0.010	0.119	0.515	0.165	0.224	0.407
	Re-ranking	BM25 + CE	$\underline{0.008}$	0.045	0.117	0.224	0.200	0.086	0.014	0.107	0.400	0.225	0.215	0.341
Symptom title + responses	Dense	all-mpnet	0.018	0.126	0.291	0.486	0.453	0.156	0.028	0.231	0.725	0.473	0.466	0.625
		multi-qa-mpnet	0.010	0.069	0.187	0.352	0.339	0.134	0.017	0.169	0.614	0.357	0.348	0.512
		DPR	0.016	0.057	0.110	0.443	0.261	0.089	0.021	0.115	0.394	0.487	0.301	0.358
		ANCE	0.021	$\underline{0.150}$	0.175	0.481	0.326	0.118	0.026	0.159	0.538	0.479	0.356	0.472
		TAS-b	$\underline{0.028}$	$\underline{0.150}$	$\underline{0.329}$	$\underline{0.590}$	$\underline{0.477}$	$\underline{0.161}$	$\underline{0.034}$	$\underline{0.233}$	$\underline{0.736}$	$\underline{0.588}$	$\underline{0.504}$	$\underline{0.650}$
	Lexical	Anserini BM25	0.022	0.118	0.228	0.548	0.413	0.121	0.031	0.218	0.564	0.542	0.445	0.518
	Sparse	DeepCT	0.005	0.037	0.108	0.200	0.199	0.085	0.012	0.106	0.395	0.203	0.201	0.328
		SPARTA	0.009	0.051	0.160	0.319	0.265	0.117	0.018	0.136	0.523	0.290	0.270	0.435
		docT5query	0.023	0.116	0.207	0.562	0.408	0.102	0.032	0.217	0.486	0.556	0.440	0.466
	Late interact.	ColBERT	0.005	0.037	0.130	0.176	0.233	0.112	0.010	0.119	0.512	0.165	0.224	0.408
	Re-ranking	BM25 + CE	0.030	0.149	0.265	0.667	0.424	0.127	0.037	0.222	0.593	0.668	0.448	0.577
Average of symptom responses	Dense	all-mpnet	0.033	0.173	**0.347**	0.692	0.512	$\underline{0.160}$	0.038	**0.261**	**0.748**	0.708	0.558	**0.674**

Table 6 (continued)

Query	Retriever	Model	AP			P			R			NDCG		
			@10	@100	@1000	@10	@100	@1000	@10	@100	@1000	@10	@100	@1000
		multi-qa-mpnet	0.034	0.165	0.322	0.691	0.486	0.153	0.039	0.250	0.717	0.717	0.536	0.649
		DPR	0.016	0.063	0.125	0.446	0.274	0.094	0.021	0.124	0.424	0.476	0.310	0.379
		ANCE	**0.035**	**0.178**	0.340	**0.721**	**0.516**	0.153	**0.040**	0.256	0.708	**0.746**	**0.565**	0.650
		TAS-b	0.022	0.107	0.227	0.489	0.386	0.134	0.028	0.190	0.627	0.500	0.414	0.543
	Lexical	Anserini BM25	0.011	0.045	0.107	0.295	0.230	0.100	0.017	0.121	0.476	0.314	0.250	0.391
	Sparse	DeepCT	0.011	0.043	0.098	0.340	0.234	0.083	0.017	0.114	0.384	0.350	0.258	0.337
		SPARTA	0.004	0.019	0.052	0.160	0.139	0.068	0.007	0.065	0.314	0.153	0.143	0.247
		docT5query	0.011	0.040	0.083	0.302	0.214	0.065	0.017	0.111	0.322	0.310	0.234	0.291
	Late interact.	ColBERT	0.033	0.153	0.311	0.673	0.463	0.151	0.038	0.234	0.698	0.686	0.511	0.628
	Re-ranking	BM25 + CE	0.029	0.083	0.145	0.619	0.229	0.100	0.035	0.121	0.476	0.635	0.311	0.429

The best performers for each block are underlined. In bold, we mark the best performers overall

- *Re-ranking model.* This search alternative enhances the initial results obtained by BM25. We employed cross-attentional re-ranking models available on the HuggingFace model hub, specifically a MiniLM cross-encoder model trained through knowledge distillation from various BERT and ALBERT variants (Wang et al., 2020).

For all these models, we maintained the default settings defined in BEIR. Further information can be found in the original manuscript (Thakur et al., 2021). The results, presented in Table 6, provide a report of performance across query variants and models. Some important observations derive from this evaluation:

- *Difficulty of the task.* Even the best models with the most consistent query formulation strategies lead to effectiveness results that have much room for improvement. Nearly all performance figures are below .700. This demonstrates that our notion of relevance, which requires: (i) topicality *and* (ii) evidence about the individual's condition in relation to the symptom, is challenging. These initial experiments open the door to further improvements in this type of sentence-based evidence detection systems.
- *Query types.* There is here a clear trend across all models and effectiveness metrics. The shortest queries (symptom title) represent the worst performing choice, while running response-based queries leads to the best results. The longest queries (symptom title concatenated with all responses) fall something in between: they are much better than symptom-only queries but fare behind response-based queries. The search task aims at locating BDI-II related sentences in the corpus and, thus, it makes sense that the best query formulation strategy stems from the use of individual textual descriptions of the BDI-II responses. Each BDI-II response exemplifies a canonical example of target expression and running a query based on this text can potentially find similar descriptions of symptoms. Note also that concatenating all responses leads to a noisier description and this manifests in the weaker results obtained for the long queries.
- *Strongest Models.* Dense retrievers stand out as the most robust search choice. In nearly all metrics and query types, the best-performing search system comes from the dense retrievers' block. This suggests that the granularity of the task (sentence-based) and the intricate notion of relevance demand a sophisticated representation of queries and sentences. This outcome contrasts with the results obtained in the original BEIR comparison (Thakur et al., 2021), where re-ranking and late-interaction models were the best zero-shot performers and BM25 was somehow competitive. Observe that our experiments are also zero-shot (no specific training made to any of the models) but we find that dense retrievers are clearly superior to the other models. This is an interesting outcome, as these dense retrievers are computationally less demanding than re-ranking models and late interaction models. Among the dense retrieval models, ANCE, TAS-b and the sentence similarity models are the most solid choices. ANCE is the best model for the optimal query formulation strategy (average of symptom's responses). However, TAS-b is superior to ANCE with the weaker set of queries (symptom title or symptom title+responses). The sentence similarity models,

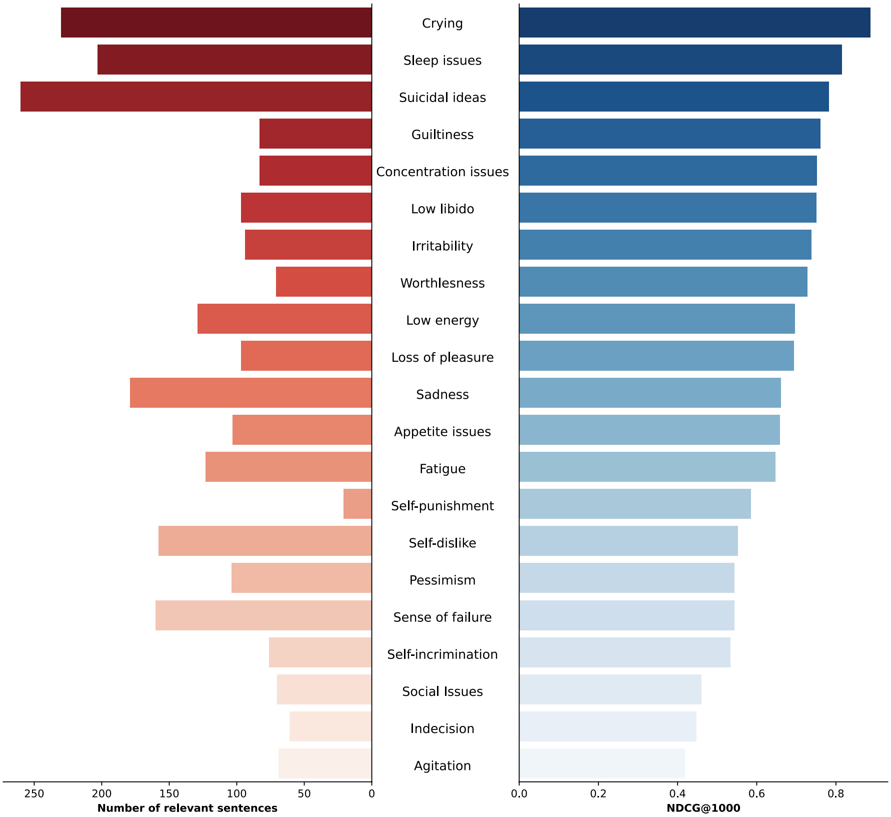

Fig. 1 Number of relevant sentences (left bars) and retrieval performance (right bars, NDCG@1000) for each BDI-II symptom. The results are those obtained with the ANCE model and the *average of symptom responses* query type

all-mpnet and *multi-qa-mpnet*, are also quite competitive in most of the cases. Note also that these two models are symmetric search models that were trained with pairs of texts of equivalent length. This explains their good performance in the third block, where queries are single sentences from individual BDI-II responses.

- *Weaker Models*. Sparse, lexical, late-interaction and re-ranking models do not seem to be competitive for this task. Among the sparse models, there is no clear winner (for example, SPARTA is the best sparse model with very short queries but it is the worst sparse model with queries based on symptom's responses). With very short queries, sparse models slightly close the gap with dense models. Sparse models work by expanding the original sentences, thereby increasing the likelihood of matches between query words and sentence words. This seems to be beneficial for symptom title queries, where query terms are scarce. On the other hand, the lexical model, BM25, only gives acceptable performance with very long queries. Still, sparse models are only clearly superior to BM25 when the queries are notably short. ColBERT, the late-interaction model, is inferior to the sparse models with very short or long queries, but it is superior to the sparse models when queries are constructed from individual BDI-II responses. Finally, the re-ranking model works in general very poorly. This could be attributed to the noise present in the original ranking. With the BDI-II item as query, BM25+CE yields better results, but only for high-precision metrics.

These patterns accentuate the necessity of model selection tailored to the specific objectives of the retrieval task. In contexts where precision is more important, models such as ANCE, from the dense retriever category, might be a good choice. Instead, if we are interested in high recall (e.g., to get all sentences related to a given BDI-II symptom) then *all-mpnet* would be a good candidate (highest AP@1000, P@1000, R@1000 and NDCG@1000).

Let us now analyse the difficulty of searching for evidence of specific BDI-II symptoms. In Fig. 1, we present the symptom-by-symptom performance of the *ANCE* model. We chose this model as an illustration because it is highly effective across most retrieval metrics, particularly for the most robust query construction strategy (individual symptom responses).

The results reveal substantial variance in detecting different symptoms. This outcome is similar to that of prior studies that examined predictive models at the symptom level (Pérez et al., 2022; Nguyen et al., 2022b; Zhang et al., 2022a). Notably, ANCE achieved high effectiveness in identifying symptoms like *Crying* (0.89) and *Changes in sleeping pattern* (0.81). In contrast, it faced difficulties with symptoms such as *Agitation* (0.42) and *Indecisiveness* (0.45). In general, the more relevant sentences available in the corpus, the higher search effectiveness. BDI-II symptoms such as *social issues*, *indecisiveness* and *agitation* are barely discussed in the collection and, thus, the search system can hardly find substantial evidence. In contrast, *crying*, *sleeping issues* and *suicidal ideas* are more recurrent topics and, as a consequence, the retrieval system does a better job at locating relevant evidence. However, there are a few BDI-II topics that do not fit with this general pattern. For example, *sense of failure* is among the most discussed topics but, still, ANCE faced difficulties to locate its relevant sentences. On the other hand, *guiltiness* or *concentration issues* do not have many relevant sentences but the system did a reasonable job at finding them. Another interesting symptom is *self punishment*. It has very few relevant sentences but ANCE was somehow effective at extracting them. In the future, we will further analyse these topics and their descriptive texts. Indeed, the peculiar characteristics of the target expressions (e.g., certain BDI-II responses) might impose difficulties or ease the extraction of relevant symptoms.

In line with findings from recent literature, performance variability may be largely attributed to the exposure of each symptom on social media. The specific nature and societal stigma associated with each symptom influences how, and if, people publicly disclose their feelings in social media writings (Nguyen et al., 2022b; Pérez et al., 2022; Zhang et al., 2022a). This symptom-level analysis underscores the need for incorporating a wide array of diverse and comprehensive training examples.

This will be specially crucial for less commonly represented symptoms, in order to enhance the models' overall effectiveness.

7 Ethics and Limitations

The primary motivation for creating DepreSym is to aid in the advancement of new technologies that can detect early indicators of depression. As pointed out in Neuman et al. (2012), it is crucial not to perceive these innovative methods and resources as substitutes for professional judgement. Automated screening methods should be regarded as digital aids that can reduce the workload on public health systems.

DepreSym can be obtained upon request, and we have adhered to strict ethical standards, particularly those pertaining to the ethical development of AI. This study did not engage with social media users directly, e.g. offering health advice. Rather, it was an observational study on open available data from Reddit. We have ensured that the data collection process preserves anonymity and rely on the exempt status under title 45 CFR §46.104. This research study was exempt from IRB review because we only experimented with existing publicly available data and we did not contact any social media user. Our primary goal is to positively impact society, with a strong focus on enhancing comprehension of depression. Making individual diagnosis of mental health is well beyond the scope of this research. It is also critical to acknowledge that the DepreSym dataset might reflect biases and limitations that are commonly found in social media studies. For example, the publications in the corpus cannot be considered a representative sample of the entire world population. Instead, we expect the presence of geographical, gender, age, or sexual orientation biases. Important segments of the population cannot be monitored from online interactions (e.g., elderly individuals who have not exposure to online platforms or people who keep their user accounts private). Furthermore, this study was confined to publications written in English. In the future, we will work towards extending this research to other languages and social media platforms, and we will try to design effective ways to understand and mitigate biases. Researchers or practitioners who utilise our data should be aware of these potential biases and take appropriate actions to adjust for fairness.

8 Conclusions

In this paper, we presented DepreSym, a novel resource aimed at advancing research in new depression screening models that leverage symptom markers at the sentence level. The sentences in this dataset were annotated through a pooling approach involving multiple ranking systems. This was complemented by a thorough assessment method involving domain experts and conversational agents. The experiments with the LLMs have manifested both their potential and limitations in this domain. While these models demonstrate a promising capacity to identify relevant sentences, their tendency towards false positives suggests a need for further refinement. This study has not only highlighted the capabilities of LLMs in understanding and categorising complex mental health data, but has also underscored the importance of ongoing development and evaluation. As we look to the future, the incorporation of other LLMs, such as LLaMA, and the implementation of hybrid approaches that combine the strengths of humans and conversational agents, hold significant promise. This approach could lead to more robust, accurate, and efficient systems for mental health screening and research.

Further, in our exploration of retrieval baselines applied to DepreSym, we assessed a variety of state-of-the-art models, including lexical, sparse, dense, and re-ranking systems. This evaluation, conducted within the BEIR framework, revealed significant insights into the dataset's inherent complexity and the challenges faced by general retrieval models in this specialized context. Notably, we observed marked improvements in retrieval effectiveness with detailed query formulations, particularly with individual symptom descriptions, highlighting the critical role of query specificity in retrieval tasks. Our analysis also included a symptom-by-symptom comparison of the performance of our methods depending on the depressive symptom considered.

Author Contributions The first and second authors contributed to the following aspects: Conceptualization, Implementation, Investigation, Writing, and Supervision. The third and fourth authors contributed to conceptualization and supervision.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. The first and third authors has received support from projects: PLEC2021-007662 (MCIN/AEI/10.13039/501100011033 Ministerio de Ciencia e Innovación, European Union Next Generation EU/PRTR) and also PID2022-137061OB-C21 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación). We also thank the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G/01, ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center. The second and fourth authors thank the financial support supplied by MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, supported by ERDF) and Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, supported by ERDF).

Data Availability DepreSym can be obtained upon request, and we have adhered to strict ethical standards, particularly those about the ethical development of AI.

Declarations

Conflict of interest We declare that we have no conflict of interest as defined by Springer or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

Alonso, O., & Mizzaro, S. (2009). Can we get rid of TREC assessors? Using mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 workshop on the future of IR evaluation* (Vol. 15, p. 16). Boston, Massachusetts, USA.

Babakov, N., Logacheva, V., & Panchenko, A. (2023). Beyond plain toxic: Building datasets for detection of flammable topics and inappropriate statements. *Language Resources and Evaluation*, 1–46.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of beck depression inventories-IA and-II in psychiatric outpatients. *Journal of Personality Assessment, 67*(3), 588–597.

Berger, A., Caruana, R., Cohn, D., Freitag, D., & Mittal, V. (2020). Bridging the lexical chasm, statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 192–199). Athens, Greece.

Büttcher, S., Clarke, C. L., Yeung, P. C., & Soboroff, I. (2007). Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 63–70). Amsterdam, Netherlands,

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine, 3*(1), 43.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality.

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights, 10*, 1178222618792860.

Crestani, F., Losada, D. E., & Parapar, J. (2022). *Early detection of mental health disorders by social media monitoring: The first five years of the ERisk project* (Vol. 1018). Springer.

Dai, Z., & Callan, J. (2020). Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1533–1536). Online Event.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Faggioli, G., Dietz, L., Clarke, C. L., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., & Wachsmuth, H. (2023). Perspectives on large language models for relevance judgment. arXiv preprint arXiv:2304.09161.

Forbes. (2022). Introducing chatgpt, November https://openai.com/blog/chatgpt. [Accessed April 4, 2023].

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056.

Harrigian, K., Aguirre, C., & Dredze, M. (2020). Do models of mental health based on social media data generalize? In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 3774–3788). Online Event.

He, X., Lin, Z., Gong, Y., Jin, A., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., & Chen, W. (2023). Annollm: Making large language models to be better crowdsourced annotators. arXiv preprint arXiv:2303.16854.

Hofstätter, Lin, S. C., Yang, J. H., Lin, J., & Hanbury, A. (2021). Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 113–122). Online Event.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906.

Khattab, O., & Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over Bert. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 39–48). Xi'an, China.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Losada, D. E., Crestani, F., & Parapar, J. (2018). Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In *Proceedings of the 9th international conference of the CLEF association, CLEF 2018, Avignon, France* (pp. 1–20).

Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR meets multilinguality, multimodality, and interaction: 10th international* *conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, proceedings 10* (pp. 340–357). Springer.

Losada, D. E., Crestani, F., & Parapar, J. (2020). Erisk 2020: Self-harm and depression challenges. In *42nd European conference on information retrieval, ECIR 2020, Lisbon, Portugal* (pp. 557–563). Springer.

Losada, D. E., & Gamallo, P. (2020). Evaluating and improving lexical resources for detecting signs of depression in text. *Language Resources and Evaluation, 54*, 1–24.

MacAvaney, S., Desmet, B., Cohan, A., Soldaini, L., Yates, A., Zirikly, A., & Goharian, N. (2018). RSDD-time: Temporal annotation of self-reported mental health diagnoses. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: From keyboard to clinic, New Orleans, LA* (pp. 168–173). Association for Computational Linguistics.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282.

Meyer, S., Elsweiler, D., Ludwig, B., Fernandez-Pichel, M., & Losada, D. E. (2022). Do we still need human assessors? Prompt-based Gpt-3 user simulation in conversational AI. In *Proceedings of the 4th conference on conversational user interfaces, Glasgow, Scotland* (pp. 1–6).

Moghadasi, S. I., Ravana, S. D., & Raman, S. N. (2013). Low-cost evaluation techniques for information retrieval systems: A review. *Journal of Informetrics, 7*(2), 301–312.

Neuman, Y., Cohen, Y., Assaf, D., & Kedma, G. (2012). Proactive screening for depression through metaphorical and automatic text analysis. *Artificial Intelligence in Medicine,56*(1), 19–25. ISSN 0933-3657.

Nguyen, T., Yates, A., Zirikly, A., Desmet, B., & Cohan, A. (2022a). Improving the generalizability of depression detection by leveraging clinical questionnaires. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics, ACL 2022, Dublin, Ireland* (pp. 8446–8459). Association for Computational Linguistics.

Nguyen, T., Yates, A., Zirikly, A., Desmet, B., & Cohan, A. (2022b). Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th annual meeting of the association for computational linguistics, Dublin, Ireland* (pp. 8446–8459). Association for Computational Linguistics.

Nogueira, R., Lin, J., & Epistemic, A. I. (2019). From doc2query to docttttquery. *Online Preprint,6*, 2.

Nuckols, C. C., & Nuckols, C. C. (2013). *The diagnostic and statistical manual of mental disorders, (DSM-5)*. American Psychiatric Association.

OpenAI. (2023). GPT-4 technical report. arXiv:submit/4812508.

Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2021). Overview of eRisk 2021: Early risk prediction on the internet. In *International conference of the cross-language evaluation forum for European languages* (pp. 324–344). Springer.

Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2023). Overview of erisk 2023: Early risk prediction on the internet. In *International conference of the cross-language evaluation forum for European languages, Thessaloniki, Greece* (pp. 294–315). Springer.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. Bloomsbury Press.

Pérez, A., Parapar, J., & Barreiro, Á. (2022). Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine, 132*, 102380. ISSN 0933-3657.

Pérez, A., Parapar, J., Barreiro, Á., & Lopez-Larrosa, S. (2023) Bdi-sen: A sentence dataset for clinical symptoms of depression. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, Taipei, Taiwan* (pp. 2996–3006). Association for Computing Machinery. ISBN 9781450394086.

Pérez, A., Warikoo, N., Wang, K., Parapar, J., & Gurevych, I. (2023). Semantic similarity models for depression severity estimation. arXiv preprint arXiv:2211.07624 To appear in: EMNLP.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing, EMNLP 2019, Hong Kong*. Association for Computational Linguistics.

Ríssola, E. A., Losada, D. E., & Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare, 2*(2). ISSN 2691-1957.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. *NIST Special Publication Sp, 109*, 109.

Sakai, T. (2009). On the robustness of information retrieval metrics to biased relevance assessments. *Journal of Information Processing, 17*, 156–166.

Santos, W. R. D., de Oliveira, R. L., & Paraboni, I. (2023). Setembrobr: A social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation,* 1–28.

Scholer, F., Turpin, A., & Sanderson, M. (2011). Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval, Beijing, China* (pp. 1063–1072).

Steer, R. A., Beck, A. T., & Garrison, B. (1986). Applications of the beck depression inventory. In *Assessment of depression* (pp. 123–142). Springer.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems, NeurIPS 2017* (Vol. 30).

Walsh, C. G., Chaudhry, B., Dua, P., Goodman, K. W., Kaplan, B., Kavuluru, R., Solomonides, A., & Subbian, V. (2020). Stigma, biomarkers, and algorithmic bias: Recommendations for precision behavioral health with artificial intelligence. *JAMIA Open, 3*(1), 9–15.

Wang, W., Wei, F., Li, D., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in neural information processing systems, NeurIPS, 2020* (Vol. 33, pp. 5776–5788).

Wolohan, J. T., Hiraga, M., Mukherjee, A., Sayyed, Z. A., & Millard, M. (2018). Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the first international workshop on language cognition and computational models, New Mexico, USA* (pp. 11–21).

Xiong, L., Xiong, C., Li, Y., Tang, K. F., Liu, J., Bennett, P., Ahmed, J., & Overwijk, A. (2020). Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv preprint arXiv:2007.00808.

Yang, P., Fang, H., & Lin, J. (2017). Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, Tokyo, Japan* (pp. 1253–1256).

Yilmaz, E., Aslam, J. A., & Robertson, S. (2008). A new rank correlation coefficient for information retrieval. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, & M.-K. Leong (Eds.), *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2008, Singapore, Singapore* (pp. 587–594). ACM.

Zhang, Z., Chen, S., Wu, M., & Zhu, K. (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, Abu Dhabi, United Arab Emirates* (pp. 9970–9985). Association for Computational Linguistics.

Zhang, Z., Chen, S., Wu, M., & Zhu, K. Q. (2022b). Symptom identification for interpretable detection of multiple mental disorders on social media. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing, EMNLP 2022, Abu Dhabi, United Arab Emirates* (pp. 9970–9985). Association for Computational Linguistics.

Zhang, Z., Chen, S., Wu, M., & Zhu, K. Q. (2022c). Psychiatric scale guided risky post screening for early detection of depression. In L. De Raedt (Ed.), *Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI 2022, Vienna, Austria* (pp. 5220–5226).

Zhao, T., Lu, X., & Lee, K. (2020). Sparta: Efficient open-domain question answering via sparse transformer matching retrieval. arXiv preprint arXiv:2009.13013.

Authors and Affiliations

Anxo Pérez[1] · Marcos Fernández-Pichel[2] · Javier Parapar[1] · David E. Losada[2]

✉ Anxo Pérez
anxo.pvila@udc.es

Marcos Fernández-Pichel
marcosfernandez.pichel@usc.es

Javier Parapar
javier.parapar@udc.es

David E. Losada
david.losada@usc.es

[1] Information Retrieval Lab, Centro de Investigación en Tecnoloxías da Información e da Comunicación (CITIC), Universidade da Coruña, Campus Elviña, 15071 A Coruña, Galicia, Spain

[2] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Galicia, Spain