Online Social Networks and Media 50 (2025) 100338

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Online expressions, offline struggles: Using social media to identify depression-related symptoms

Mario Ezra Aragón [a,*], Adrián Pastor López-Monroy [b,c,*], Manuel Montes-y-Gómez [d], David E. Losada [a]

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Rúa Jenaro de la Fuente, 15782, A Coruña, Spain
[b] Centro de Investigación en Matemáticas (CIMAT), A.C., Jalisco S/N, Col. Valenciana, 36023, Guanajuato, Mexico
[c] Universidad Virtual del Estado de Guanajuato (UVEG), Hermenegildo Bustos #129 A Sur, Col. Centro, 36400, Guanajuato, Mexico
[d] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro #1, 72840, Puebla, Mexico

ARTICLE INFO

Keywords:
Depression
Social media
Beck's Depression Inventory
Large language models

ABSTRACT

With their growing popularity, social media platforms have become valuable tools for researchers and health professionals, offering new opportunities to identify linguistic patterns associated with mental health. In this study, we analyze depression-related symptoms using user-generated posts on social media and the Beck Depression Inventory (BDI). Using posts from individuals who have self-reported a depression diagnosis, we train and evaluate sentence classification models to assess their ability to detect BDI symptoms. Specifically, we conduct binary classification experiments to identify the presence of depression-related symptoms and additional tests to categorize sentences into specific BDI symptom types. We also perform a comprehensive symptom-level analysis to examine how depressive symptoms are expressed linguistically, linking social media data with a clinically validated framework. In addition, we analyze symptom distributions between users with and without depression and across platforms, providing insight into how symptoms manifest in diverse online contexts. Furthermore, we incorporate a data augmentation strategy that leverages Large Language Models to generate clinically grounded synthetic examples and evaluate their effectiveness against human-generated data. Our findings indicate that users with depression exhibit a significantly higher prevalence of certain BDI symptoms – particularly Suicidal Thoughts, Crying, Self-Dislike, and Changes in Sleeping Pattern – while control users predominantly express milder categories such as Sadness or Pessimism. Synthetic data improves the detection of underrepresented symptoms and enhances model robustness, although human-generated data better captures subtle linguistic nuances. Specialized models outperform general ones, but specific symptom categories remain challenging, underscoring the need for more interpretable and clinically grounded detection frameworks.

1. Introduction

Mental health conditions affect countless individuals worldwide, disrupting their thoughts, behavior, and overall quality of life [1]. Among psychological disorders, depression is one of the most common and is a leading contributor to suicidal behavior, making it a critical public health concern [2]. Major depressive disorder (MDD) is a severe medical condition that negatively impacts emotions, thoughts, and daily activities. Individuals with MDD often experience persistent feelings of sadness, hopelessness, emptiness, and a marked loss of interest or pleasure in activities they once found enjoyable. These emotional challenges frequently extend into physical symptoms, such as chronic fatigue or unexplained aches. These psychological barriers can disrupt relationships, careers, and everyday responsibilities [3]. Common symptoms of depression include diminished interest in routine activities, an increase in feelings of guilt or worthlessness, notable changes in weight or appetite, and disturbances in sleep patterns — ranging from insomnia to excessive sleeping. If left untreated, these symptoms can escalate, leading to severe impairment in an individual's ability to function socially and professionally. Currently, only about 20% of individuals affected by depression receive the necessary early intervention, and a significant portion of mental health funding is directed toward maintaining psychiatric institutions rather than focusing on detection, prevention, and recovery efforts [4]. This highlights an urgent need to develop effective strategies for the early detection of

* Corresponding authors.
E-mail addresses: ezra.aragon@usc.es (M.E. Aragón), pastor.lopez@cimat.mx (A.P. López-Monroy).

https://doi.org/10.1016/j.osnem.2025.100338
Received 1 July 2025; Received in revised form 8 October 2025; Accepted 8 October 2025
Available online 14 October 2025

depression, aiming to mitigate harm and improve outcomes for those living with this condition.

Given the profound impact of mental health disorders, it is crucial to thoroughly understand their symptoms, severity, and progression over time. This understanding is crucial in advancing our knowledge of these conditions and developing effective models to support mental health practitioners. By identifying patterns and changes in symptoms, practitioners can provide more personalized and timely interventions, ultimately improving the support available to individuals experiencing psychological challenges.

The impact of new mental health screening forms goes beyond individual care. Computer-based technologies that monitor online evidence provide valuable insights into broader mental health trends, enabling the identification of at-risk populations and informing public health initiatives. By addressing shared symptoms and experiences, new prevention strategies and early intervention programs can be developed. This can benefit many individuals who may not yet be officially diagnosed but are exhibiting signs of mental health concerns.

We live in a digital era where user-generated content is published on social media platforms daily, providing researchers with unique opportunities to examine how individuals confront different challenges. Social media has become a space where humans openly share their daily experiences, significant life events, and personal reflections. For some, these platforms offer a sense of community and a safe space, mainly through anonymity, enabling them to discuss sensitive topics like mental health without fear of judgment [5,6]. This openness has created a rich environment for studying how individuals express their emotions and manage mental health concerns, thus allowing researchers to gain valuable insights about depression.

Most existing studies on depression detection primarily focus on binary user-level classification, aiming to distinguish between individuals with and without depression [5]. While effective in broad categorization, these approaches often lack interpretability and fail to identify explicit, standardized symptomatology that could be meaningfully integrated into clinical practice. This limitation significantly hinders their applicability in healthcare settings, where a detailed understanding of symptom patterns is essential for accurate diagnosis, informed treatment decisions, and effective communication between clinicians and patients. Consequently, there is a critical need for models that move beyond simple classification to provide symptom-level insights aligned with established diagnostic criteria. In 2019, eRisk [7] launched a pioneering project to assess the feasibility of automated methods to estimate the intensity of specific symptoms of depression. Unlike traditional approaches, this initiative targeted a fine-grained analysis of standardized clinical indicators derived from the Beck Depression Inventory [8].

The BDI is a structured psychological inventory. It provides a standardized and objective approach to evaluating specific aspects of human behavior [9]. The BDI serves as a comprehensive framework for measuring various symptoms associated with depression, making it a valuable resource for research and clinical practice. Designed as a self-assessment tool, the BDI evaluates typical attitudes and symptoms linked to depression. Through 21 multiple-choice questions, it measures the intensity and prevalence of emotions such as sadness, hopelessness, self-disapproval, indecisiveness, and low energy levels. The BDI principle is that negative cognitive distortions play a crucial role in the development and persistence of depression [8].

In this study, we build on existing research to analyze depression-related symptoms by examining user-generated content on social media platforms and leveraging the BDI questionnaire to track signs of risk. By integrating social media and clinical knowledge, we aim to gain a deeper understanding of the connection between online behavior and specific clinical symptoms of depression. To achieve this, we employed state-of-the-art classification technology to detect depression symptoms in social media posts, addressing two main challenges:

1. **Symptom detection:** Determine whether a sentence reflects any of the symptoms of depression.
2. **BDI symptom classification:** Focuses on fine-grained detection of specific symptoms, building 21 independent (binary) classifiers or, alternatively, a single multiclass classifier encompassing 22 classes (21 corresponding to the BDI questions and one class for non-symptoms).

Our goal is to investigate how effectively symptoms can be identified in online content and to what extent computational tools based on clinical questionnaires can help recognize patterns of depression. The main focus is on three research questions (RQ1–RQ3), which address computational modeling and data augmentation. We also consider two exploratory questions (RQ4 and RQ5) on sociological and cross-platform aspects. These latter questions are intended as preliminary analyses to illustrate possible extensions of our framework, with a deeper treatment left for future work. These research questions are as follows:

- **RQ1:** Are current text classification models effective in categorizing depression symptoms? Do mental health-adapted models outperform more general language models?
- **RQ2:** Which BDI symptoms are more difficult to detect in social media environments, and what are the possible barriers (e.g., data scarcity, symptoms expressed in subtle ways, subjectivity, vocabulary diversity)?
- **RQ3:** How do Large Language Models perform in generating synthetic BDI-related examples to train classifiers?
- **RQ4:** How different are the distributions of depression symptoms between users who have been diagnosed with depression and those who have not been diagnosed? Are the automated methods detecting more BDI-related pieces of evidence among users suffering from depression?
- **RQ5:** How does the manifestation of depression symptoms on social networks vary across different platforms?

This study aims to inform the design of novel risk assessment tools that support decision-making and mental health interventions, while laying the groundwork for more extensive cross-platform and sociological analyses in future research. We summarize our main contributions as follows:

1. A comprehensive analysis of the presence of specific depression symptoms on social media. In this study, we leverage the Beck Depression Inventory questionnaire as a reference for identifying and classifying symptoms in user posts.
2. A detailed study on how synthetically generated data helps in detecting symptoms of depression, comparing its effectiveness with human-generated sentences.
3. An in-depth exploration of how depression symptoms manifest in users with and without depression. This analysis is complemented by exploring these trends on different social media platforms.

The remainder of the paper is organized as follows. Section 2 reviews existing literature on detecting signs of depression through social media data. Section 3 outlines the main steps of our methodology, providing detailed technical descriptions. Section 4 provides a detailed description of the experimental setup, while Section 5 presents a thorough analysis of our experiments and results, highlighting the key findings. Section 6 offers an evaluation and analysis of the incorporation of augmented data. Section 7 presents our study of depression symptoms across various social media platforms and user types. Section 8 discusses the limitations and ethical considerations of our work. Finally, Section 9 concludes the paper, summarizing the main insights derived from our study.

2. Related work

Depression is a mental health disorder characterized by a persistent loss of interest in daily activities, leading to significant challenges in everyday life. The worldwide mental health crisis has hindered access to mental healthcare, resulting in many disorders going untreated [10]. Many studies focused on the automatic detection of depression, often utilizing information from social media [5]. One common approach consists of extracting features from users' social media posts and building predictive models based on the extracted features [11]. Some strategies relied on linguistic features, such as word frequencies and n-grams, to encode language patterns [12–15]. These models typically employed traditional classification algorithms like support vector machines, random forests, or neural networks. For example, in [16], the authors developed a generalized machine learning approach for detecting depression in social media texts, even when explicit keywords (e.g., 'depression' or 'diagnosis') are absent. The study explored various text preprocessing methods, textual-based features, and machine learning classifiers (both single and ensemble models). However, distinguishing users with depression from those with no depression can be challenging due to substantial overlap in vocabulary usage. More recently, the rise of transformer-based models, such as BERT, has offered new possibilities. BERT-based classifiers have been fine-tuned to detect signs of various mental health conditions, including depression [17–19]. These models generally outperform traditional approaches in classification tasks, providing a more effective means of identifying psychological risks [20].

Recent advances in automated depression detection from social media have leveraged deep learning and large language models (LLMs) to enhance predictive accuracy and interpretability. Shah et al. [21] demonstrated that fine-tuned LLMs, specifically GPT-3.5 Turbo and LLaMA2-7B, can achieve high accuracy in detecting depressive content. Similarly, Figuerêdo et al. [22] proposed a CNN-based model enhanced with fusion strategies and emoticon-based emotion mapping to improve early depression detection. Their approach revealed that the semantic interpretation of emoticons boosts precision and recall, highlighting the importance of affective signals in user-generated content. Pradnyana et al. [23] introduced an ensemble model that integrates personality traits and sentiment polarity patterns to improve performance and explainability. Their hybrid system – combining RoBERTa with RF-BiLSTM – demonstrated increased accuracy and F1-score over existing models. Ilias and Askounis [24] explored the relationship between stress and depression through a multi-task learning (MTL) paradigm. By jointly modeling these two interlinked conditions using shared and task-specific BERT layers, they achieved improved performance over single-task and transfer learning baselines.

Various studies have employed self-reported inventories, such as the Beck Depression Inventory, to identify social media users potentially experiencing depression [25]. For instance, De Choudhury et al. [26] pioneered automated depression detection by gathering data on depression prevalence using crowd-sourcing techniques. These authors applied the CES-D (Center for Epidemiologic Studies Depression Scale) [27] to measure the participants' depression levels and obtained permission to access their public Twitter feeds. In another study [28], Skaik and Inkpen utilized the eRisk 2021 data [29] to develop an automated system capable of estimating responses to the BDI questionnaire provided by Reddit users. These authors created specific models for each BDI question and integrated the best-performing models into a single framework called the "BDI-Multi-Model". This integrated approach surpassed the performance of previous state-of-the-art solutions for this challenging task. Additionally, Kang et al. [30] utilized an advanced deep-learning model to predict BDI scores based on electroencephalogram (EEG) data. Their findings confirmed the potential for accurately forecasting BDI scores using EEG signals.

Although a few studies have explored the identification of specific depression symptoms through assessment questionnaires or related tools, much remains to be investigated [31]. One noteworthy example in this area is the work by [32], which introduces an unsupervised method to align content from various mental health-related subreddits with DSM-5 categories. This alignment relied on a domain-specific lexicon incorporating n-grams associated with mental health disorders described in the DSM-5. The lexicon was successfully exploited to analyze the relationship between subreddit content and DSM-5 categories. Similarly, [33] focused on creating a depression-specific lexicon that captured common symptoms of depression based on the PHQ-9 clinical assessment inventory [34]. This lexicon served as background knowledge for generating a set of seed terms for each symptom. Leveraging these seeds, the authors trained a semi-supervised topic model using tweets from users who self-identified as experiencing depression. More recently, [35] introduced KoMOS, a Korean mental health dataset annotated at the symptom level by experts, and proposed an innovative framework for detecting mental disorders by utilizing symptoms and their contextual information. This framework leverages large language models to extract and analyze context-rich data for symptom detection. In [36], Guo et al. presented a dataset from Sina Weibo and approached depression detection as a binary classification problem. Employing the Dalian University of Technology Sentiment Lexicon (DUT-SL), this team developed a specialized depression lexicon with enriched linguistic features.

Social media platforms offer textual information and rich metadata, providing valuable insights into mental health. For instance, in [37], the authors demonstrated how Twitter metadata can be an effective tool for the early detection of depression. This team developed a comprehensive system with a multi-modal and multi-task approach, where detecting depression was the primary goal and recognizing emotions was a secondary task. Building on this, Ghosh and Anwar [38] focused on predicting individuals experiencing depression and assessing its severity using Twitter data. Their self-supervised approach generated preliminary labels by extracting user characteristics, including emotional, thematic, behavioral, user-specific, and depression-related n-gram attributes. Sadasivuni et al. [39] introduced another innovative method by analyzing tweets related to depression topics and proposed a novel parameter to quantify the level of depression on any given day. Compared with historical data, this parameter becomes a valuable tool for social scientists studying the effectiveness of psychotherapy and temporal trends in depression prevalence.

In recent work [40], Belcastro and colleagues introduced an effective and interpretable approach for analyzing depression through social media posts. This research marks significant progress by utilizing interpretable AI techniques that provide clear insights alongside accurate predictions of depression. In particular, the potential integration of Large Language Models (LLMs) offers exciting possibilities for further enhancing the explainability of automated screening methods.

Overall, substantial effort has been devoted to identifying and predicting signs of depression through monitoring online sources, reflecting the growing interest in using digital platforms for depression screening. Despite these advancements, research focusing on the presence and distribution of specific depression symptoms remains comparatively limited. Moreover, there is a notable gap in understanding the textual characteristics that enable practical and accurate markers of depression. To address these gaps, our research closely examines the manifestation of depression symptoms in online publications. Specifically, we aim to explore how standardized BDI symptoms are expressed, their frequency and distribution across social media posts, and their occurrence patterns, specifically related to individuals who suffer from depression. By gaining deeper insights into these aspects, we hope to contribute to the understanding of how signs of depression are communicated online and inform the development of more effective tools for identifying and addressing mental health challenges in digital spaces.

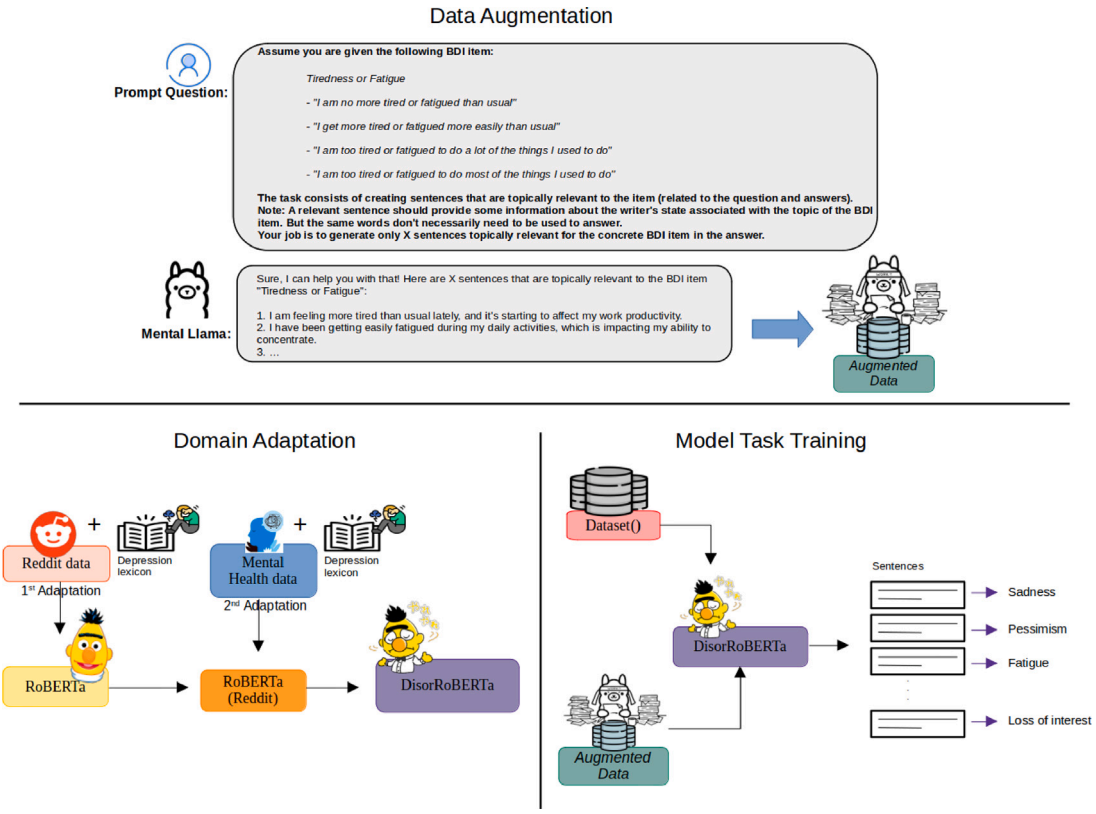

Fig. 1. General diagram. At the top, we illustrate the data augmentation strategy, which involves asking the MentalLlama model to generate relevant sentences. At the bottom, we show the domain adaptation of the language model and the downstream task (sentence classification using the original dataset and augmented sentences).

3. Methodology

This section describes our methodology for adapting language models to detect depression symptoms and the data augmentation process to generate new training samples. The approach has three main stages: data augmentation, domain adaptation, and task adaptation. During the first stage, we ask a Large Language Model to generate new samples related to the different BDI items. BDI-labeled sentences do not abound; thus, it is crucial to assess the ability of LLMs to generate synthetic sentences. In the second stage, we use a double-domain strategy to adapt a base model from a general domain (e.g., RoBERTa or BERT) to a specialized mental health domain. Finally, in the third stage, we build a sentence-level classifier using an original collection of sentences annotated for depression symptoms and the augmented data. Fig. 1 illustrates the entire process, including data augmentation, double domain adaptation, and the final training for the target task.

3.1. Data augmentation using large language models

Building large datasets, especially in specialized domains like mental health, can be expensive, time-consuming, or ethically challenging. These challenges become particularly pronounced when humans are needed for annotation purposes. While human labeling is valuable for creating high-quality datasets, it poses difficulties arising from human limitations, effort, task complexity, and logistic constraints. Therefore, generating synthetic samples becomes a practical alternative for enriching existing datasets. In our work, a data augmentation procedure is applied to expand the eRisk dataset (see Section 4.1 for a detailed description), but the synthetic BDI-related sentences generated could also be exploited to enrich other corpora of sentences (e.g., collections labeled in terms of psychological dimensions).

We begin the process by generating new samples for each symptom of the BDI, namely (1) Sadness, (2) Pessimism, (3) Past Failure, (4) Loss of Pleasure, (5) Guilty Feelings, (6) Punishment Feelings, (7) Self-Dislike, (8) Self-Criticalness, (9) Suicidal Thoughts, (10) Crying, (11) Agitation, (12) Loss of Interest, (13) Indecisiveness, (14) Worthlessness, (15) Loss of Energy, (16) Changes in Sleeping Pattern, (17) Irritability, (18) Changes in Appetite, (19) Concentration Difficulty, (20) Tiredness or Fatigue, and (21) Loss of Interest in Sex. The synthetic samples are sentences conveying BDI symptoms in different ways.

To generate new samples related to the symptoms of depression, we employ MentalLlama [41] and the Ollama framework.[1] We prompted specific instructions to the model, asking it to generate new sentences related to each BDI item. The instructions follow the assessment guidelines of the original (human-labeled) collections. For example, we explicitly indicate that a sentence deemed relevant for a BDI symptom should be on-topic and, additionally, offer some insight into the state of the person who writes the sentence. At the top of Fig. 1, we present an example of the prompt and the model's answer. The core idea is that a relevant sentence should provide evidence about an individual suffering from (or not suffering from) the BDI symptom. Note that a relevant sentence might positively or negatively inform about the individual's state. For example, for the tiredness/fatigue symptom, "I never get fatigued" and "I am feeling more tired than usual lately" are both relevant because these two texts are informative about the tiredness/fatigue state of the person who wrote this sentence.

[1] https://github.com/ollama/ollama

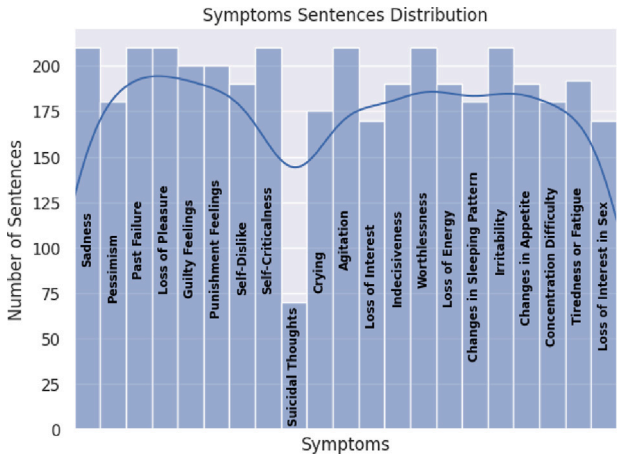

Fig. 2. Distribution of synthetic sentences generated.

Table 1
Augmented data statistics.

number of sentences	3947
average length (# words)	14.95
max length (words)	33
min length (words)	7
vocabulary (# words)	879

We prompted the model to generate 210 examples for each symptom. Still, depending on the symptom, the model ended up generating fewer examples, especially for sentences related to suicidal thoughts (BDI item 9). This is somehow expected since LLMs have ethical limitations on generating text about specific topics. In the current implementation, we did not include real social media posts as examples within the prompts used for data augmentation (i.e., we did not implement few-shot). This was a deliberate design choice aimed at maintaining generalization across symptoms and avoiding overfitting to the specific linguistic or topical patterns of the dataset available. Our prompts merely described the semantic and emotional characteristics of each BDI symptom in natural language, based on the eRisk annotation guidelines. While this approach ensured consistency and interpretability, we acknowledge that incorporating paraphrased or anonymized examples from real posts could enhance the realism and lexical diversity of the generated data. We highlight this few-shot alternative as a promising direction for future work. Fig. 2 presents the distribution of synthetic sentences generated. Table 1 presents additional statistics of the augmented data, e.g., the total number of sentences, their average length, and vocabulary size.

3.2. Domain adaptation

For domain adaptation, we adopt the approach proposed in [42,43], which extends the model's pre-training through additional fine-tuning of the masked language model. This fine-tuning consists of two stages, where the first exploits a general Reddit corpus and the second a collection of documents related to mental health. This dual process adapts the original model from the general language of Wikipedia and book corpora to the more specific language of Reddit and, then, to the specifics of mental health-related texts. This training process is essential for optimizing the model's performance in downstream classification tasks related to psychological analysis in social media.

Following the methodology [44], we implemented a two-stage domain adaptation process using either RoBERTa or BERT as base models. In brief, the approach first teaches RoBERTa/BERT with language sourced from social media (Reddit) and then further specializes the model with expressions from users discussing topics related to mental disorders.

The first adaptation stage adjusts the model to the language style prevalent in social media. To that end, we used the corpus from [45], which contains pre-processed Reddit posts designed for text summarization. For the second adaptation, we used datasets from subreddits focused on mental health and depression.[2] By concentrating on these topics, the model learns to identify how users with mental disorders express themselves and the specific language they use on social media.

To improve the learning process, we incorporated additional lexical knowledge related to depression. BERT family models randomly mask training words to (self-)train the language model. This involves selecting a random number of words in a sentence and asking the model to predict the hidden word. In this way, the model learns different contexts where each word occurs. In our approach, we implemented an enhanced variant of masking supported by a depression lexicon. Instead of randomly masking words, masking was steered toward depression words: if a sentence contained words from the depression lexicon, then these words were masked; if the masked words did not reach 20% of the total words in the sentence, we added random words to complete the required masking. This modified approach helps the model focus more on words related to mental disorders. The depression lexicon we used is from [46], one of the few publicly available resources specifically targeting depression.

We concatenated all examples from the mental health datasets and divided the resulting corpus into equal-sized chunks of 128 words each. Then, we masked 20% of the words in each batch. We trained the model over three epochs during both stages of domain adaptation with a batch size of 128 and a learning rate of $2e^{-5}$. We utilized an NVIDIA Tesla V100 32 GB SXM2 GPU to perform the domain adaptation and fine-tuning. We elaborate on the details of each step below and refer to the final models as DisorROBERTa or DisorBERT, respectively.

3.3. Downstream task adaptation

As mentioned above, our primary goal is to develop a deeper understanding of how online behavior is linked to specific clinical symptoms of depression. We address different classification challenges to extract traces of symptoms. To achieve this, we built multiple models designed to detect symptoms of depression in sentences from social media posts. We adopted two distinct strategies:

Symptom Detection: This strategy predicts whether a sentence conveys any of the symptoms from the BDI questionnaire (regardless of the specific symptom). Under this *binary* classification approach, we categorize each sentence as positive if relevant to any depression symptom and negative otherwise. The primary objective is to assess the effectiveness of various models in identifying markers associated with depression.

BDI Symptom Classification: This fine-grained strategy identifies specific symptoms as represented in the BDI inventory. To that end, we experiment with two alternative methods. The first method builds 21 independent BDI-item classifiers, where each model detects a specific symptom from the BDI questionnaire. Each model is trained using a 1 vs. all approach (the examples from a given item are taken as positive samples, and all examples from the rest of the classes act as negative examples, including those from the non-symptom class). The second method involves the construction of a single *multiclass* classifier, with 22 classes (one class for each BDI item and another class representing "no symptom"). This approach aims to assess the models' ability to differentiate between various symptoms of depression.

Furthermore, we integrated the augmented data into both strategies and evaluated the performance of the models with and without synthetic examples. By incorporating additional data, we aimed to enhance the models' ability to detect symptoms related to depression.

[2] https://huggingface.co/datasets/ShreyaR/DepressionDetection, https://huggingface.co/datasets/hugginglearners/reddit-depression-cleaned, and https://huggingface.co/datasets/jsfactory/mental_health_reddit_posts

4. Experimental settings

4.1. Dataset

For the evaluation, we utilized the datasets provided by the eRisk 2023 and 2024 evaluation initiatives [47,48]. These shared-data evaluation campaigns were designed to encourage advancements in identifying early indicators of mental health risks on social media. Over the years, the eRisk tasks have addressed early risk challenges for diverse psychological disorders, including anorexia, depression, gambling, and self-harm. The evolution of the eRisk tasks reflects the growing importance of designing proactive approaches that monitor mental health in social media spaces [6]. A key innovation introduced for the 2023 eRisk campaign [47] was a new challenge focused on finding sentence-level evidence for each BDI symptom. Specifically, participants were asked to rank sentences (drawn from user-generated content) according to their estimated relevance to symptoms from the Beck Depression Inventory-II (BDI-II) questionnaire. A sentence was considered relevant if it provided positive or negative evidence about the user's state concerning a specific symptom. It is essential to note that positive sentences can be considered relevant to a given symptom. For instance, a statement such as *"I feel quite happy lately"* would be regarded as relevant to the symptom *"Sadness"* as it directly reflects the user's emotional state concerning that dimension of depression.

The original datasets provided to the eRisk's participants were formatted according to the TREC standards and consisted of a corpus of sentences derived from eRisk's previous datasets (social media publications). Three expert assessors were asked to annotate a pool of sentences associated with each symptom to obtain relevance judgments. The assessors were given detailed instructions to help them determine relevance. The dual requirement for relevant sentences, topical relevance, plus reflection of the user's state added a layer of complexity beyond traditional relevance assessments. To manage this challenge, the assessors developed a robust annotation methodology supported by formal assessment guidelines, ensuring consistency and accuracy in their judgments. We recommend consulting the task's description paper [47] for a more detailed explanation of the dataset creation process and the criteria for selecting relevant sentences.

Although the dataset was initially created for evaluating search models, we utilize it here for classification purposes. To that end, we extracted each sentence annotated by the assessors and, based on its assigned relevance, assigned a label of 1 to indicate relevance or 0 otherwise (spanning categories 1 through 21, corresponding to the 21 BDI items). Table 2 provides a comprehensive summary of the main statistics of the resulting dataset.

Table 2
Depression symptom distribution in the eRisk dataset, where #sen represents the number of sentences, %sen the percentage of sentences, avglen the average length of the sentences (# words), and vocab the number of unique words (excluding stopwords) in the sentences from the category.

#	Label	#sen	%sen	avglen	vocab
0	No Symptom	17,414	63.14%	20.03	10,473
1	Sadness	684	2.48%	9.53	354
2	Pessimism	567	2.06%	16.64	621
3	Past Failure	468	1.70%	12.28	433
4	Loss of Pleasure	319	1.16%	14.49	401
5	Guilty Feelings	509	1.85%	10.38	420
6	Punishment Feelings	212	0.77%	12.00	247
7	Self-Dislike	570	2.07%	9.98	404
8	Self-Criticalness	391	1.42%	12.66	397
9	Suicidal Thoughts	722	2.62%	11.26	451
10	Crying	729	2.64%	8.89	429
11	Agitation	530	1.92%	12.59	570
12	Loss of Interest	299	1.08%	14.61	395
13	Indecisiveness	443	1.61%	11.56	348
14	Worthlessness	321	1.16%	8.02	223
15	Loss of Energy	377	1.37%	12.29	404
16	Changes in Sleeping	689	2.50%	16.31	601
17	Irritability	441	1.60%	11.59	459
18	Changes in Appetite	543	1.97%	14.59	626
19	Concentration Difficulty	415	1.50%	13.11	471
20	Tiredness or Fatigue	479	1.74%	11.98	471
21	Loss of interest in Sex	456	1.65%	17.12	525

4.2. Training configurations

Preprocessing: We began with basic text preprocessing, converting all words to lowercase and removing elements such as URLs, emoticons, and hashtags. This step was designed to eliminate non-essential components that do not contribute meaningfully to our analysis.

Training and predictions: To determine an appropriate input length, we experimented with truncation thresholds of 50, 70, and 100 tokens. Model performance plateaued at 70 tokens, offering a practical balance between context retention and computational efficiency. Importantly, the eRisk dataset predominantly contains short user-generated sentences, with an average length of approximately 20 tokens and a maximum relevant length of around 60. As such, truncation at 70 tokens preserved the full content of most inputs and had minimal impact on semantic completeness or symptom expression. For the symptom detection task, each sentence was labeled as positive when deemed relevant for any of the 21 symptoms and negative otherwise. Consequently, the positive set contains the union of relevant sentences from the 21 BDI symptoms. In constructing 21 independent BDI classifiers, we labeled a sentence as positive if it was relevant for the specific symptom and negative otherwise. Lastly, for the multiclass approach (22-class), we assigned to each sentence a symptom (if it was relevant to that symptom) or None (if it was non-relevant to all BDI symptoms).[3] Finally, we applied a 5-fold cross-validation for symptom detection and a 3-fold cross-validation for the two other fine-grained experiments.

Parameters: For data augmentation, we used the default Ollama implementation of MentalLlama,[4] with temperature set to 0.7, top-p to 0.9, and top-k to 40. For each prompt, a single response was generated under these fixed parameters, as we aimed to obtain consistent outputs for evaluation rather than to compare across multiple generations. No additional filtering or post-processing of responses was applied. To assess robustness, we manually inspected representative outputs and observed that repeated runs with the same prompt and parameters yielded highly similar generations, indicating stable model behavior under the chosen settings.

For domain adaptation, we employed the frameworks offered by HuggingFace v4.24.0 [43] and PyTorch v1.13.0 [49] for model development. Specifically, during the training phase, we adopted a batch size of 256, Adam optimizer with a learning rate set to $1e^{-5}$, and cross-entropy as the loss function. The training process spanned three epochs.

For the classification experiments, we also utilized the frameworks provided by HuggingFace (v4.24.0) and PyTorch (v1.13.0) for the downstream task's training and conducted the training over three epochs. We used different hardware configurations for different stages of our pipeline based on the computational demands of each task. Domain adaptation was performed on an NVIDIA V100 (32 GB) GPU due to its higher memory requirements, particularly when fine-tuning large language models with massive datasets. For classification experiments, which involved smaller variants and fewer parameters, we used a GeForce RTX 3070 (8 GB) GPU.

[3] In the resulting corpus, there are no sentences with multiple BDI labels; thus, we treat the problem as a single label-multiclass problem.

[4] https://ollama.com/vitorcalvi/mentallama2

4.3. Classification models

GPT-4 0-shot: This unsupervised approach utilizes GPT-4 under a zero-shot setting. To that end, we prompt the model to assign the symptom label that best describes the sentence. Since it is a non-supervised method, it ignores the information available in the training splits. This unsupervised baseline helps contextualize the results of the other supervised methods.

Bag-of-Words (BoW): We implemented a traditional Bag-of-Words approach using word unigrams weighted by TF-IDF. These classical representations were then classified by a Support Vector Machine (SVM) with a linear kernel. We made preliminary tests with alternative kernels and other traditional classification methods, but the linear SVM consistently yielded the best performance among these classic alternatives. We therefore adopted it as our reference traditional baseline.

BERT: We utilized a BERT-based (uncased[5]) classification model, applying fine-tuning to adapt the pre-trained model to each specific training set.

RoBERTa: Similar to BERT, we employed a RoBERTa-based[6] classification model and fine-tuned it on each training set. The fine-tuning process helped RoBERTa leverage its robust pre-training on large datasets while tailoring it to our specific context.

DisorBERT: DisorBERT [44] is a customized variant of BERT that undergoes a dual-domain adaptation process. First, the model was adapted to social media language to capture informal and conversational textual patterns. Next, it specializes in the mental health domain, enhancing its ability to recognize domain-specific terminology and linguistic patterns related to psychological disorders. Both adaptation steps used a domain-specific lexicon to guide the masking process, ensuring the model focuses on relevant language features during training.

DisorRoBERTa: Similar to DisorBERT, this represents our double-domain adaptation of a RoBERTa language model.

GPT-4 embeddings: We extracted sentence embeddings generated by GPT-4 and employed an SVM with a linear kernel as the classifier. This combination allowed us to evaluate GPT-4's capabilities in capturing meaningful text representations for classification.

Llama3 embeddings: We used sentence embeddings obtained from the Llama3.1 model via the Ollama platform. Similar to the previous case, these embeddings were fed into an SVM with a linear kernel for classification.

MentalLlama embeddings: MentalLlama [41] is a specialized variation of the Llama model. It was explicitly trained on mental health-related data to enhance its understanding of this domain. Using the Ollama platform, we extracted sentence embeddings from MentalLlama and applied a linear SVM classifier.

5. Evaluation results

In this section, to answer the questions posed by RQ1 –Are current text classification models effective in categorizing depression symptoms? Do mental health-adapted models outperform more general language models?–, we report and analyze the results of the proposed models for the two tasks. The first series of experiments did not incorporate data augmentation. The discussion on the effect of data augmentation is deferred until Section 6.

[5] https://huggingface.co/google-bert/bert-base-uncased
[6] https://huggingface.co/FacebookAI/roberta-base

Table 3
Symptom detection task (binary classification). F1, Precision (P), and Recall (R) results for the positive class.

Model	P	R	F1
GPT-4 (0-shot)	0.534 ± 0.192	0.821 ± 0.071	0.631 ± 0.142
BoW	0.645 ± 0.134	0.756 ± 0.070	0.693 ± 0.102
BERT	0.668 ± 0.147	0.889 ± 0.023	0.756 ± 0.095
RoBERTa	0.665 ± 0.149	0.914 ± 0.027	0.763 ± 0.103
GPT-4	**0.712 ± 0.109**	0.826 ± 0.073	0.763 ± 0.089
Llama3	0.631 ± 0.191	0.734 ± 0.149	0.653 ± 0.109
MentalLlama	0.702 ± 0.126	0.719 ± 0.124	0.703 ± 0.102
DisorBERT	0.684 ± 0.135	0.898 ± 0.038	0.772 ± 0.094
DisorRoBERTa	0.679 ± 0.137	**0.916 ± 0.040**	**0.774 ± 0.097**

5.1. Symptom detection

Table 3 summarizes the results obtained by the different methods on the binary classification task of symptom detection (symptom vs no symptom). The evaluation metrics include precision, recall, and F_1 score, focusing on the positive class (i.e., depression-related symptoms). Notably, the GPT-4 0-shot variant (first row) does not perform task-based training, whereas all other models are supervised and thus exploit the designated training split. To ensure robust evaluation, we employed a five-fold cross-validation approach and reported the average performance and standard deviations across the folds. This experimental setup enables us to evaluate the consistency and reliability of the models across various subsets of the data.

As a first observation, we emphasize the improved performance of the adapted models compared to their base counterparts. This outcome shows the value of domain-specific models in addressing specialized tasks. DisorRoBERTa outperformed the baseline models regarding F_1 and recall, demonstrating its ability to retrieve a higher proportion of relevant results. On the other hand, the model based on GPT-4's embeddings yielded the highest precision, followed by MentalLlama's embeddings. The MentalLlama model outperformed the Llama3 model, offering a more balanced trade-off between precision and recall. This reinforces the benefit of incorporating domain-specific information.

The binary classification results indicate robust performance in risk detection, effectively identifying multiple indicators of psychological risks. This capability is critical in clinical screening, where high recall is essential for comprehensive evidence extraction. However, it is also important to consider scenarios where prioritizing precision may be more appropriate. For example, in tracking social media to identify and address the riskiest behaviors, a precision-oriented approach could help minimize false positives and focus interventions where they are most needed. Further model adaptations to enhance precision could be critical in such contexts, opening future research opportunities.

5.2. Individual BDI symptom classifiers

Fig. 3 plots the F1 performance achieved by each classification model for the 21 BDI categorization tasks. Overall, we can highlight that DisorRoBERTa achieves the highest mean F1 performance, 0.737. The DisorBERT model comes in second, with an average of 0.724, showing good performance across most BDI items. The GPT-4 0-shot model yields significantly weaker results than fine-tuned models, highlighting the importance of task-specific training.

According to our experiments, the symptoms easier to detect are: "Loss of energy", "Concentration difficulty", "Guilty feelings", "Crying", and "Changes in appetite". The classification models effectively capture the textual nuances associated with these symptoms. In contrast, symptoms such as "self-criticalness", "Sadness", or "Past Failure" are hard to detect (even the best-performing models struggle, with modest F1 scores of 0.677). For "changes in sleeping pattern", DisorRoBERTa performed the best (0.726), but other models, including

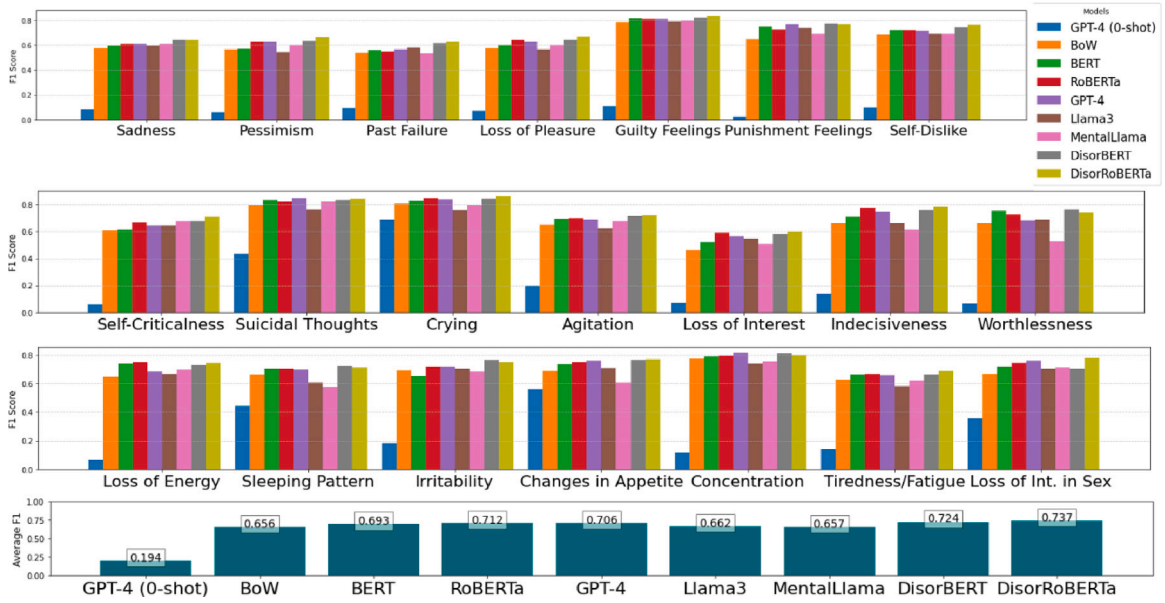

Fig. 3. Classification results for the 21 BDI symptom classifiers.

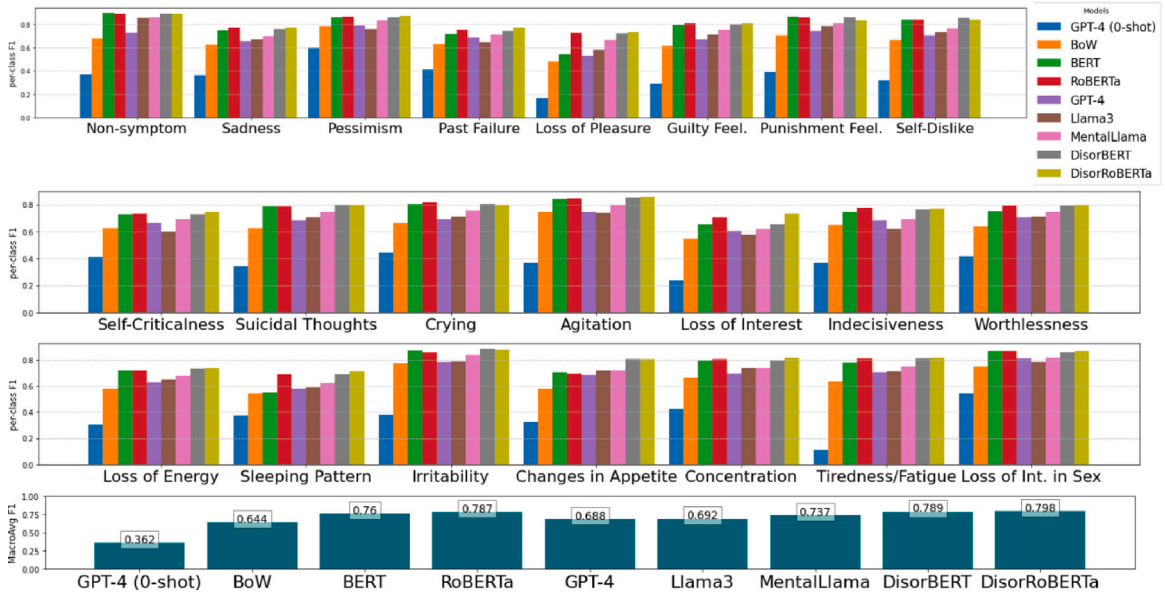

Fig. 4. 22-class classification results: macro-average F1 (bottom plot) and per-class F1 (1st, 2nd and 3rd plots).

Llama3 and GPT-4, performed poorly. This suggests that capturing subtle textual signals related to some BDI symptoms is challenging. Symptoms like "Suicidal Thoughts", "Crying", and "Agitation" show wider performance gaps among the models, with DisorRoBERTa again outperforming other classifiers.

It is essential to notice that embedding-based models like MentalLlama and GPT-4 consistently outperform the zero-shot variant of GPT-4. This difference highlights the superiority of fine-tuning models based on embedded representations compared with other variants without task supervision.

In conclusion, DisorRoBERTa is the strongest performer across nearly all symptoms, achieving the highest average F1 score. Other models like DisorBERT and RoBERTa also excel, while general-purpose or zero-shot models perform significantly worse.

5.3. 22-Class classifier

We now measure the capability of the models in the multi-class setting (22 classes: 21 BDI items + no BDI category). Fig. 4 presents the results of the different models (including the macro-averaged F1 and the per-class F1 values). We can observe that, again, DisorRoBERTa achieves the highest global average F1 score at 0.798, followed closely by DisorBERT with 0.789. Both models consistently outperform other methods, showing their robustness for the task.

Among the non-adapted models, RoBERTa stands out, demonstrating its strong effectiveness as a pre-trained model. MentalLlama also performs well, with a global average of 0.737; GPT-4 (0-shot) performs poorly, with a global average of 0.362, suggesting that zero-shot models

struggle in this 22-class classification task. GPT-4 performs better than GPT-4 (0-shot) (0.688 vs 0.362) but still lags behind RoBERTa-based methods.

Analyzing the effectiveness for specific symptoms, we can observe that symptoms such as "Pessimism", "Agitation", or "Loss of Interest in Sex" achieve high F1 scores across all models, particularly with DisorRoBERTa and DisorBERT, exceeding 0.85 in most cases. This indicates that these classes are easier to identify, possibly due to more straightforward language patterns or well-defined linguistic cues.

In contrast, challenging symptoms like "Loss of Interest", "Changes in Sleeping Pattern", and "Changes in Appetite" yield low F1 scores, especially with Llama3 and GPT-based methods. This suggests that these classes are more ambiguous or complex.

Some symptoms present a wide variability in results. For example, "Suicidal Thoughts" and "Self-Criticalness" show significant performance differences among models, with DisorBERT and DisorRoBERTa significantly outperforming baseline models.

Embedding-based models (e.g., MentalLlama and GPT-4) generally outperform zero-shot methods, reinforcing the importance of fine-tuning or task-specific supervision. Traditional models like Bag-of-Words (BoW) remain the least competitive, with a global average of 0.644.

5.4. Error analysis

To address our second research question –"Which BDI symptoms are more difficult to detect in social media environments, and what are the possible factors (e.g., data scarcity, symptoms expressed in subtle ways, subjectivity, vocabulary diversity)?"– we conducted a detailed analysis at the symptom level of the best-performing model, DisorRoBERTa. Fig. 5 shows a confusion matrix of the model's predictions for the 22-class classifier, offering a comprehensive overview of the errors made.[7]

Symptoms like "Crying", "Suicidal Thoughts", "Guilt Feelings", and "Loss of interest in Sex" have the best-performing results, with values near or equal to 0.93 or 0.97 on the diagonal. This indicates a high precision in the model's predictions for these categories. Notably, for example, "Worthlessness" and "Loss of Energy" also have relatively high accuracy, around 0.89 and 0.88, respectively.

Conversely, symptoms such as "Loss of Pleasure" and "Loss of Interest" show a noticeable number of off-diagonal values. For instance, "Loss of Pleasure" is sometimes confused with "Loss of Interest", suggesting the model may struggle to differentiate between these two closely related symptoms. Similarly, "Sadness" is somewhat confused with "Pessimism" or "Crying", indicating overlap or similarity in how these are interpreted or labeled. Another example is "Past Failure", which is commonly misclassified as "Self-dislike", suggesting again potential ambiguities between these related symptoms.

We can highlight the following areas for improvement: "Past Failure" has a significant misclassification rate, as do "Loss of Interest" and "Loss of Pleasure". These values suggest potential areas where the model can improve through better feature extraction or more distinctive training data.

"Tiredness or Fatigue" and "Changes in Sleeping Pattern" show some degree of overlap, suggesting that fatigue or tiredness symptoms often interact with sleeping patterns.

Overall, the model performs well, particularly on specific symptoms like "Crying" and "Suicidal Thoughts or Wishes". However, the classification performance could further benefit from strategies targeting specific confusions among overlapping categories, particularly those associated with losses, self-reflection, and energy-related symptoms.

To deepen this analysis, we examined some misclassified sentences by the model (see Table 4). First, we observe that the model struggles to distinguish between closely related emotional states and categories, resulting in frequent misclassification. Some sentences often reflect ambiguity, and potentially, multiple symptoms could have been assigned to some of these sentences.

For example, in the "Agitation" block, common misclassifications are linked to symptoms like sleeping difficulties, irritability, and fatigue. For example, "I have trouble calming down at night and falling asleep" was labeled by the model as "Changes in Sleeping Patterns", indicating difficulty distinguishing agitation from the derived sleeping problems. Arguably, this sentence is relevant to both categories; however, the collection was built using a pooling approach, and as a consequence, only the top-ranked sentences for each symptom were annotated.

The misclassified sentences for "Past Failure" are tied to self-perception and emotions such as pessimism, worthlessness, and self-dislike. Some examples are: "I feel pathetic and also convinced I actually am pathetic", which was misclassified as "Worthlessness" despite fitting into broader self-critical themes. Regarding "Loss of Interest", misclassification is often related to no symptom assignments, concentration, or loss of pleasure. For instance, "I'm bored with my life" is labeled as "Loss of Pleasure", reflecting the model's struggle to recognize the nuanced distinction between interest and pleasure.

These sentences, therefore, illustrate complex structures in which multiple symptoms may coexist. This complexity makes accurate classification more challenging and causes misclassification issues. For example, ambiguity in language, with sentences expressing overlapping symptoms, often misleads the model. The subjective nature of the task and how individuals self-report their symptoms complicate the assignment of BDI categories. Furthermore, the training data is imbalanced, as specific categories are underrepresented, which affects performance.

These results present opportunities for improvement. For example, we could refine the model to handle nuanced language more effectively (e.g., by incorporating additional context or augmenting the training datasets with more diverse examples for specific symptoms). In the following section, we explore this second possibility.

6. Data augmentation analysis

Data augmentation is a technique used in machine learning and data science to artificially expand the size and diversity of a dataset by applying various transformations to the original data [50–52]. By creating variations of the existing data or new synthetic examples, data augmentation helps models generalize better, improves performance on unseen data, and mitigates overfitting. It is especially beneficial when collecting additional real-world data is costly, time-consuming, or impractical. Nowadays, LLMs have gained popularity in the research community due to their capacity in various fields and their ability to comprehend and generate coherent text. Therefore, it is natural to consider the exploitation of LLMs for data augmentation.

This section addresses our third research question: "How do Large Language Models perform in generating synthetic BDI-related examples to train classifiers?". Specifically, we explore the ability of LLMs to generate new sentences related to the 21 BDI symptoms (Section 3.1). For this series of experiments, we begin by training the classifiers exclusively on the synthetic data. Then, we combine the generated examples with the original training data. These experiments were done with the most effective classifier (DisorROBERTa).

The first experiment used only the generated data and tested the 21 BDI classifiers when trained exclusively with synthetic data. To that end, each example was considered positive for the BDI item that generated the example and negative for the rest of the BDI items.

Fig. 6 presents the results of this experiment. As expected, the original data proved to be more effective than the artificial data. Nonetheless, the average performance of the classifier built from synthetic data alone is 0.546, surpassing the 0-shot GPT-4 model and approaching traditional methods like BoW. This highlights the potential

[7] The original counts have been normalized to proportions where the sum of each row's values is equal to 1.

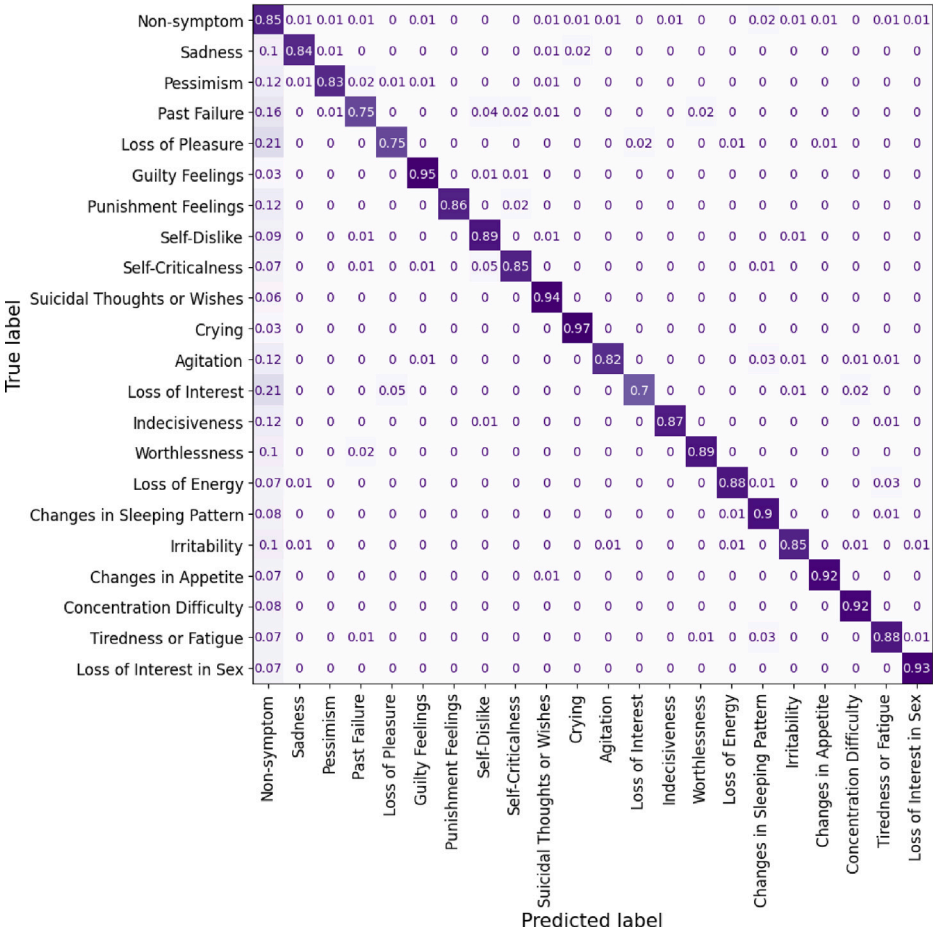

Fig. 5. Confusion Matrix of the DisorRoBERTa model at symptom level (22-class).

Table 4
Examples of sentences misclassified by the classification model (DisorRoBERTa). In the first column, we show examples of three true categories (Agitation, Past Failure, and Loss of interest), and the second column indicates the (wrong) label assigned by the model.

Agitation	Predicted as
I have trouble calming down at night and falling asleep.	Sleeping...
Anxiety kind of forces me to multitask and complete things fully and quickly.	Concentration
I don't know how people co-sleep, I'm too anxious.	Sleeping...
I am exhausted and I don't feel safe falling asleep.	Sleeping...
It was a restless sleep, and I ended up sleeping in the passenger seat.	Sleeping...
But I can't seem to do so – I just end up feeling guilty, restless, and trapped.	Guilty
I feel like the room is closing in on me, I panic, and everything is extremely irritating to me.	Irritability
I'm usually stressed out, and tired by nightfall.	Fatigue
I just feel really depressed and anxious and I wanna sleep but I'm not tired so I can't force myself.	Fatigue
Past Failure	
My depression has made me apathetic as a result of my apathy I am going to fail an apprenticeship I have spent the last year doing.	Pessimism
I feel like a waste of space, a burden, and just someone who makes everyone's life worse.	Worthlessness
I feel pathetic and am also convinced I actually am pathetic as that's how society judges anyone without a social life.	Worthlessness
I'm proud of myself lmao	Self-Dislike
I know I dwell in the past a lot and beat myself up over not doing things sooner.	Self-Criticalness
I have not been proud of myself in several years.	Self-Dislike
I don't resent myself, I've learned that every mistake I've made has put me where I'm at now and I'm happy and content.	Self-Dislike
Loss of interest	
I have no excitement or thrill of getting to know people.	Loss of Pleasure
I'm bored with my life	Loss of Pleasure
I love to talk and meet new people.	Non-symptom
I have my interests and what I'm looking for, in not much detail.	Non-symptom
My interests are varied but tbh I can't focus on much.	Concentration
I can't focus, I've lost interest in almost everything, and I've lost almost all self-confidence.	Concentration
I'm not as passionate as I used to be, not as outgoing.	Non-symptom

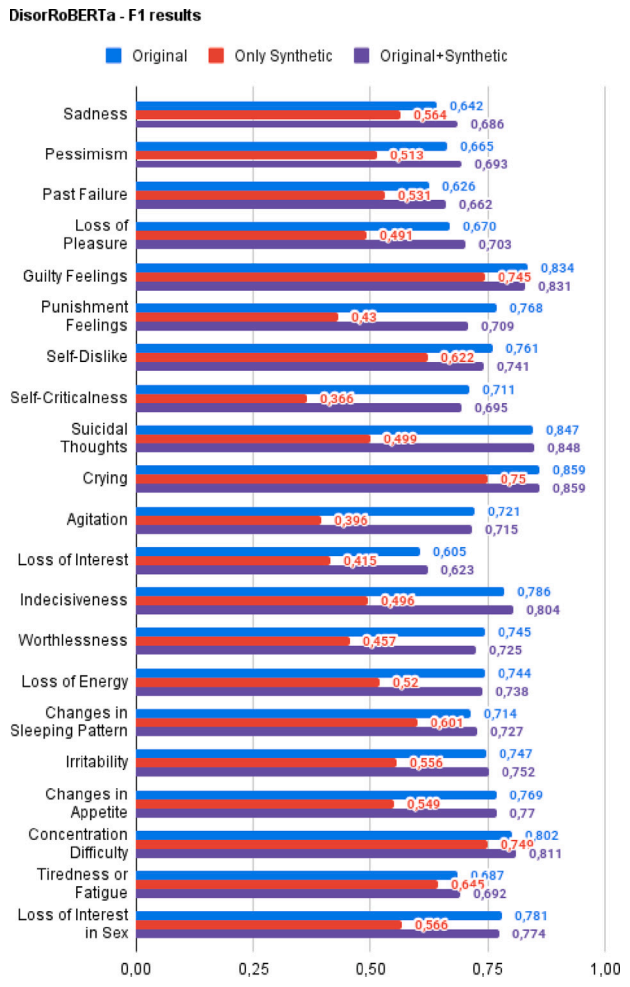

Fig. 6. F1 results using the original data, the synthetic data, and the combination of both.

of LLMs in generating BDI-related sentences, saving time and human effort in labeling high volumes of data. The combination of original and synthetic outperforms or closely matches the original performance across all categories. Incorporating synthetic examples led to substantial improvements over the original F1 for some categories. For example, adding synthetic examples to the original data made the BDI classifier of "Sadness" go from 0.642 to 0.686, or the BDI classifier of "Pessimism" go from 0.665 to 0.693.

Our next experiment evaluated the effect of synthetic data on the symptom detection task (binary). We compare the original classifier with three data augmentation methods, based on adding all generated examples or only examples generated for five or seven low-performing BDI items.[8] This targeted augmentation aimed to alleviate specific model predictions' weaknesses. The results of this experiment are summarized in Table 5. The model yields high recall, demonstrating its ability to identify the most positive instances. However, this comes at the expense of precision, as all variants show a higher rate of false positives. The F1 score reflects a reasonable balance between precision and recall but leaves room for improvement. The augmented data consistently improved the model's precision; specifically, the approach incorporating examples only from five low-performing BDI items offered the best outcome.

[8] The 5 or 7 BDI categories with the lowest F1 performance in the original experiments (please refer to Fig. 5).

Table 5
Symptom Detection Task (binary classification). F1, Precision (P), and Recall (R) results of the original model and three augmented variants.

Model	P	R	F1
DisorRoBERTa	.679 ± .137	**.916 ± .040**	.774 ± .097
DisorRoBERTa+Augment(all)	.680 ± .136	.915 ± .040	.776 ± .097
DisorRoBERTa+Augment(7)	.689 ± .128	.907 ± .044	.779 ± .093
DisorRoBERTa+Augment(5)	**.690 ± .128**	.912 ± .043	**.781 ± .091**

Table 6
22-class classifier. F1 macro average of the original model and three augmented variants.

Model	F1 macro
DisorRoBERTa	0.798
DisorRoBERTa+Augment(all)	0.796
DisorRoBERTa+Augment(7)	0.802
DisorRoBERTa+Augment(5)	**0.806**

Next, we analyze the effect of augmented data on the 22-class classifier. Similar to the previous experiment, we compare the incorporation of synthetic examples for all categories with two variants incorporating smaller sets of synthetic examples (from the 5 or 7 low-performing symptoms). Table 6 reports the macro average results. Again, including synthetic examples for low-performing symptoms led to the best-performing variants. However, the improvements are modest compared to those in the symptom detection task.

In summary, targeting specific BDI strategies not handled well by the original classifiers seems promising, and a careful selection of augmentations can refine the model's output. These initial results open an exciting opportunity to continue exploring these techniques in future research.

To contextualize our findings, we compare them now with some prior studies. Traditional machine learning approaches using n-gram features and ensemble classifiers achieved modest performance levels (ranging from .5 to .6 F1 depending on the test collection) for depression detection [14,15], and lacked symptom-level granularity and interpretability. More recent methods tackled symptom classification directly and reported improved results using specialized models [25]. Our approach builds on this by integrating domain-adapted models and data augmentation, improving classification task settings, and outperforming the symptom-level baselines reported. Additionally, while other works proposed using LLMs to enhance performance [21], their results were not focused on fine-grained BDI classification. Compared with these studies, our method performs competitively, specifically targeting symptom-level detection. This comparison highlights the novelty and effectiveness of our pipeline, especially in combining clinical alignment (via BDI) with LLM-based augmentation and dual-domain adaptation.

7. Depression symptom manifestation on social media

The analysis of depression symptoms across different social media platforms provides valuable insights into how individuals express their psychological states and how they cope with mental health challenges. Each platform has distinct features, from text-focused interactions on X (formerly Twitter) or Reddit to image sharing on Instagram. These functionalities influence how symptoms such as sadness or low self-esteem are communicated. Understanding these platform-specific differences is essential for tailoring interventions and support mechanisms.

In this section, we present an exploratory analysis addressing research questions four and five: "How different are the distributions of depression symptoms between users who have been diagnosed with depression and those who have not been diagnosed? Are the automated methods detecting more BDI-related pieces of evidence among users suffering from depression?" and "How does the manifestation

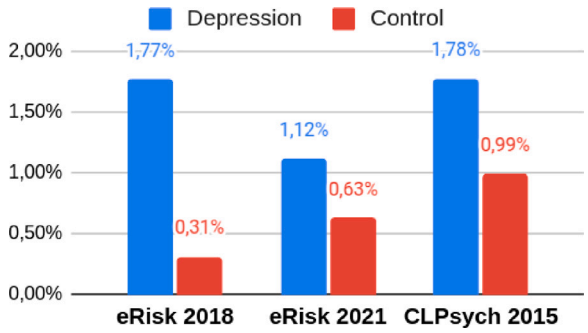

Fig. 7. Average presence of symptoms over the user's publications in the respective datasets.

of depression symptoms on social networks vary across different platforms?". To investigate these questions, we analyze datasets from the literature containing user-generated content from diagnosed (positive) and non-diagnosed (control) users. We extract all posts for each user and label them using our previously trained DisorRoBERTa model (22-class categorization). The model assigns a BDI symptom label when relevant evidence is detected or a "non-symptom" otherwise.

We worked with the collections provided by the eRisk 2018 [53] (Reddit data), eRisk 2021 [29] (Reddit data), and CLPsych 2015 [54] (X data) campaigns. These initiatives established a common evaluation framework to foster research on the automated recognition of signs of mental disorders through monitoring social media. Fig. 7 presents the average percentages of estimated symptoms in publications from depressive and control users. For the three datasets, the depressive users present a significantly higher presence of BDI symptoms in their posts. However, note that the percentages are low (below 2%), showing that most online publications do not contain explicit evidence of a BDI symptom. This is somehow expected since most users do not constantly talk about psychological issues but instead publish content about different activities and topics (expressing a wide range of emotions or no emotions at all).

We now analyze the distribution of specific depression symptoms in the users' posts. Fig. 8 plots the distribution of BDI symptoms computed over the set of posts that were assigned a BDI category (i.e., removing the "no symptom" class). For depression users (left-hand side plots), we can observe a higher prevalence of concerning symptoms such as "Suicidal Thoughts", "Crying", and "Self-Dislike". Furthermore, "Changes in Sleeping Pattern" also exhibit notable trends in the depression group. Symptoms such as "Sadness", "Pessimism", and "Crying" display a more similar distribution between depression and control users. In contrast, for the control groups (right-hand side plots), we can observe that "Sadness", "Pessimism", and "Changes in Appetite" are the most prominent symptoms across all datasets. Overall, control users not only make fewer posts showing evidence of signs of depression, but when they do express symptoms, they tend to be a much milder category. For example, we observe higher proportions of symptoms such as sadness or pessimism, while the proportion of suicidal thoughts is much lower.

Table 7 highlights significantly more prevalent symptoms on one platform than the other. Notably, users on X (formerly Twitter) exhibit a higher presence of symptoms such as crying and tiredness. In contrast, on Reddit, symptoms like suicidal thoughts, agitation, and loss of interest in sex appear more frequently. This variation may come from the differing levels of privacy offered by these platforms. Reddit, which allows for more anonymity, might encourage users to disclose more sensitive or stigmatized symptoms. In contrast, platforms like X, where users often share personal information more openly (and, typically, the account profile reveals the user's name), may lead to a different pattern of symptom expression. Furthermore, the length of the posts

Table 7
Symptom variation between Reddit and X users.

High presence in X and low in Reddit
Crying, Tiredness
High presence in Reddit and low in X
Suicidal thoughts, Agitation, Loss of interest in sex

and the type of posts could also explain the difference between the two platforms. Observe that Reddit communities are more focused, as they are groups formed to share information about specific topics. These findings can potentially inspire new lines of research exploring more specialized models for symptom detection of depression or other mental disorders.

8. Limitations and ethical considerations

Our primary objective is to advance the understanding of depression symptoms, ultimately promoting societal well-being. However, this research is not designed to diagnose mental health conditions for specific individuals.

While our study focuses on the technical feasibility of classifying BDI-related symptom expressions, we recognize the ethical and practical risks of applying such systems in real-world settings. Automated mental health screening tools may produce false positives or false negatives, which can lead to unintended consequences such as unnecessary concern, missed interventions, or even stigmatization — particularly if used without clinical oversight. Moreover, our data sources are limited to publicly available posts, written in English, and derived from Reddit. This introduces demographic and linguistic biases, potentially excluding non-English speakers, older adults, and individuals from underrepresented communities or from segments of the population that are less digitally active. As such, the generalizability of our findings is limited, and future work should address these gaps by incorporating additional datasets and ensuring that proper ethical safeguards, such as user consent, transparency, and human-in-the-loop oversight, are incorporated. Nevertheless, our analytical strategies are highly adaptable and can seamlessly apply to other languages and data sources. This flexibility enables the study of how depression symptoms manifest across different countries and cultures, allowing for a deeper understanding of the global cultural and societal factors influencing depressive behaviors.

While our approach demonstrates the potential of using LLMs to generate BDI-aligned symptom descriptions, we acknowledge that synthetic data in the mental health domain demands rigorous validation. In this study, we performed limited manual spot-checking of generated content – randomly sampling 10–15 sentences per symptom category – to ensure alignment with BDI criteria and to confirm the absence of clinically implausible or offensive outputs. No significant problems were identified in this sample, although this process was not exhaustive. For future work, we recognize the importance of incorporating systematic human-in-the-loop evaluation, ideally involving clinicians or trained mental health professionals, to ensure clinical validity and mitigate risks of hallucinated or overgeneralized symptoms. Such validation would enhance the trustworthiness and applicability of synthetic data in sensitive domains like mental health.

We acknowledge that the single-label classification setup used in our study does not fully reflect the clinical reality, where symptoms frequently co-occur and individual sentences may express multiple aspects of depression. This design choice was selected due to the structure of the eRisk dataset, in which annotations are symptom-specific and collected independently, resulting in some sentences being labeled as relevant to more than one symptom. However, such multi-label cases accounted for less than 3% of the dataset, and for consistency, we restricted our analysis to sentences with a single positive label. While

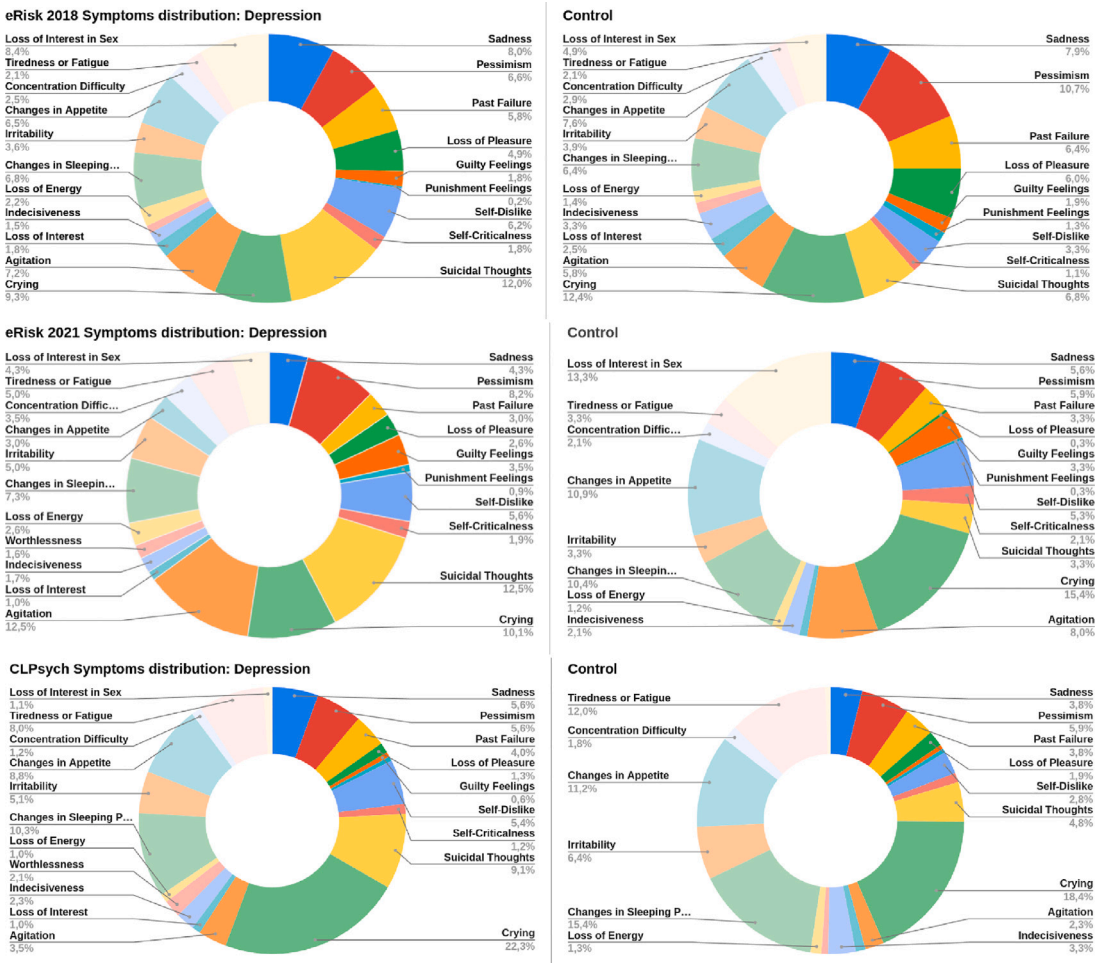

Fig. 8. BDI symptom distribution (set of posts that were assigned a BDI category) over the three datasets (eRisk 2018, eRisk 2022, and CLPsych 2015).

this simplification enables a more straightforward modeling setup, it overlooks the nuance of overlapping symptoms such as self-criticalness and worthlessness, or agitation and sleep disturbances. In this respect, multi-label modeling represents a promising direction for future work.

We recognize the ethical complexities of analyzing social media content, particularly concerning privacy issues. To address these challenges, our study relied solely on preexisting, publicly accessible datasets. We did not interact with or contact social media users at any point. The datasets included only public user interactions and were used strictly in compliance with the respective collections' terms of service and user agreements. For all datasets, user identifications (e.g., account names) were anonymized to ensure individuals' privacy and confidentiality. We only used public data and did not contact social media users, and this research project was formally approved by our IRB review (USC 65/2024).

The datasets have limited sample sizes (due to the complexities of data collection and the need to respect privacy considerations). This study took an observational approach and did not collect personal or clinical data from participants. While such data is often essential for risk assessment in clinical settings, it is beyond the scope of our research.

9. Conclusion

In this study, we investigate depression-related symptoms by analyzing user posts on social media and exploit the Beck Depression Inventory questionnaire to support symptom-level classification. Our goal is to advance automated methods for screening depression symptoms.

We conducted experiments through various classification approaches in two key scenarios: (1) to detect the presence of any BDI symptom (binary categorization) and, additionally, (2) to classify sentences into BDI-induced categories. The series of experiments was designed to compare multiple classification methods, including supervised models and 0-shot approaches.

Additionally, we implemented a data augmentation strategy using large language models to generate synthetic sentences relevant to each BDI category. Our evaluation showed the potential of LLMs to enhance the detection of specific symptoms. Additionally, we examined the distribution of symptoms between users diagnosed with depression and those without a diagnosis. The results indicate that combining domain adaptation, lexical knowledge, and data augmentation improves the detection of depression symptoms. This joint approach outperformed state-of-the-art baselines.

For future research, we plan to apply more specialized lexical resources tailored to the target tasks and incorporate clinical data to train advanced language models. Since emojis play a key role in social media analysis, we also aim to integrate them into the training process. Furthermore, we are interested in extending this study to multiple languages, as much of the research on mental disorders has predominantly focused on English.

CRediT authorship contribution statement

Mario Ezra Aragón: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Adrián Pastor López-Monroy:** Writing – review & editing, Validation. **Manuel Montes-y-Gómez:** Writing – review & editing, Supervision, Formal analysis. **David E. Losada:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank SECIHTI for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies, as well as for the support received through project CBF-2025-I-4384 of the *Ciencia Básica y de Frontera 2025* program. Mario Ezra Aragón and David E. Losada thank the support obtained from 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). The first author also thanks the support obtained from the Juan de la Cierva Grant (JDC2023-052296-I), funded by MCIN/AEI/10.13039/501100011033 and by the FSE+.

Data availability

Data will be made available on request.

References

[1] R. Kessler, E. Bromet, P. Jonge, V. Shahly, Marsha, The burden of depressive illness, Public Heal. Perspect. Depressive Disord. 4 (2017) 0–66.

[2] C.D. Mathers, D. Loncar, Projections of global mortal- ity and burden of disease from 2002 to 2030, in: PLOS Medicine, Public Library of Science, 2006, pp. 1–20.

[3] A.P.A. APA, What is depression?, 2020, https://www.psychiatry.org/patients-families/depression/what-is-depression.

[4] M.E. Renteria-Rodriguez, Salud mental en mexico. NOTA-incytu número 007, 2018.

[5] D.E. Ríssola, F. Crestani, A survey of computational methods for online mental state assessment on social media, ACM Trans. Comput. Heal. 2 (2021) http://dx.doi.org/10.1145/3437259.

[6] F. Crestani, D.E. Losada, J. Parapar, Early detection of mental health disorders by social media monitoring, in: The First Five Years of the ERisk Project, Springer Verlag, Englewood Cliffs, NJ, 2022.

[7] D.E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D.E. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, vol. 34, Springer International Publishing, Cham, 2019, pp. 0–357.

[8] A.T. Beck, R.A. Steer, M.G. Carbin, Psychometric properties of the beck depression inventory: Twenty-five years of evaluation, Clin. Psychol. Rev. 8 (1988) 77–100.

[9] S. Urbina, A. Anastasi, Psychological Testing, Seventh ed., Prentice Hall., Upper Saddle River, N.J, 1997.

[10] S.T. Ibrahim, M. Li, J. Patel, T.R. Katapally, Utilizing natural language processing for precision prevention of mental health disorders among youth: A systematic review, Comput. Biol. Med. 188 (2025) 109859, http://dx.doi.org/10.1016/j.compbiomed.2025.109859, https://www.sciencedirect.com/science/article/pii/S0010482525002094.

[11] J. Aguilera, D.I.H. Farías, R.M. Ortega-Mendoza, M.Montes-y. Gómez, Depression and anorexia detection in social media as a one-class classification problem, Appl. Intell. 51 (2021) 6088–6103, http://dx.doi.org/10.1007/s10489-020-02131-2.

[12] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, H. Ohsaki, Recognizing depression from twitter activity, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2015, pp. 3187–3196, http://dx.doi.org/10.1145/2702123.2702280.

[13] H.A. Schwartz, J. Eichstaedt, M.L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, L. Ungar, Towards assessing changes in degree of depression through facebook, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal To Clinical Reality, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 118–125, http://dx.doi.org/10.3115/v1/W14-3214, https://aclanthology.org/W14-3214.

[14] G. Coppersmith, C. Harman, M. Dredze, Measuring post traumatic stress disorder in twitter, Proc. Eighth Int. AAAI Conf. Weblogs Soc. Media 57 (2014) 9–582.

[15] G. Coopersmith, M. Dredze, C. Harman, Quantifying mental health signals in twitter, in: Workshop on Computational Linguistics and Clinical Psychology, 2014.

[16] R. Chiong, G.S. Budhi, S. Dhakal, F. Chiong, A textual-based featuring approach for depression detection using machine learning classifiers and social media texts, Comput. Biol. Med. 135 (2021) 104499, http://dx.doi.org/10.1016/j.compbiomed.2021.104499, https://www.sciencedirect.com/science/article/pii/S0010482521002936.

[17] R. Martńez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers : ilab at clef erisk 2020, in: CEUR Workshop Proceedings 2696, Thessaloniki, Greece, September 22-25, 2020, in: working Notes of CLEF 2020 - Con- ference and Labs of the Evaluation Forum, 2020, urn:nbn:de:0074-2696-0. https://strathprints.strath.ac.uk/72995/.

[18] J. Parapar, P. Martín-Rodilla, D.E. Losada, F. Crestani, Overview of erisk 2022: Early risk prediction on the internet, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2022, pp. 233–256.

[19] D. Kodati, R. Tene, Identifying suicidal emotions on social media through transformer-based deep learning, Appl. Intell. 53 (2023) 11885–11917, http://dx.doi.org/10.1007/s10489-022-04060-8.

[20] M. He, E.M. Bakker, M.S. Lew, Dpd (depression detection) net: a deep neural network for multimodal depression detection, Heal. Inf. Sci. Syst. 12 (2024) 53, http://dx.doi.org/10.1007/s13755-024-00311-9.

[21] S.M. Shah, S.A. Gillani, M.S.A. Baig, M.A. Saleem, M.H. Siddiqui, Advancing depression detection on social media platforms through fine-tuned large language models, Online Soc. Netw. Media 46 (2025) 100311, http://dx.doi.org/10.1016/j.osnem.2025.100311, https://www.sciencedirect.com/science/article/pii/S2468696425000126.

[22] J.S.L. Figuerêdo, A.L.L. Maia, R.T. Calumby, Early depression detection in social media based on deep learning and underlying emotions, Online Soc. Networks Media 31 (2022) 100225, http://dx.doi.org/10.1016/j.osnem.2022.100225, https://www.sciencedirect.com/science/article/pii/S2468696422000283.

[23] G.A. Pradnyana, W. Anggraeni, E.M. Yuniarno, M.H. Purnomo, An explainable ensemble model for revealing the level of depression in social media by considering personality traits and sentiment polarity pattern, Online Soc. Netw. Media 46 (2025) 100307, http://dx.doi.org/10.1016/j.osnem.2025.100307, https://www.sciencedirect.com/science/article/pii/S2468696425000084.

[24] L. Ilias, D. Askounis, Multitask learning for recognizing stress and depression in social media, Online Soc. Networks Media 3 (2023) 7–38, http://dx.doi.org/10.1016/j.osnem.2023.100270, 100270. https://www.sciencedirect.com/science/article/pii/S2468696423000290.

[25] E. Bao, A. Pérez, J. Parapar, Explainable depression symptom detection in social media, Heal. Inf. Sci. Syst. 12 (2024) 47, http://dx.doi.org/10.1007/s13755-024-00303-9.

[26] M.D. Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, USA, 2013.

[27] L.S. Radloff, The ces-d scale: A self-report depression scale for research in the general population, Appl. Psychol. Meas. 1 (1977) 385–401.

[28] R.S. Skaik, D. Inkpen, Predicting depression in Canada by automatic filling of Beck's depression inventory questionnaire, IEEE Access 10 (2022) 102033–102047.

[29] J. Parapar, P. Martín-Rodilla, D.E. Losada, F. Crestani, Overview of erisk 2021: Early risk prediction on the internet, in: K.S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, 32, Springer International Publishing, Cham, 2021, pp. 4–344.

[30] M. Kang, S. Oh, K. Oh, S. Kang, Y. Lee, The deep learning method for predict beck's depression inventory score using eeg, in: 2021 International Conference on Information and Communication Technology Convergence, ICTC, vol. 49, 2021, pp. 0–493.

[31] A.P. Association, Diagnostic and Statistical Manual of Mental Disorders, fifth ed., American Psychiatric Publishing, Washington, 2013.

[32] M. Gaur, U. Kursuncu, A. Alambo, A. Sheth, R. Daniulaityte, K. Thirunarayan, J. Pathak, Let me tell you about your mental health!: Contextualized classification of reddit posts to dsm-5 for web-based intervention, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, vol. 75, 2018, pp. 3–762.

[33] A.H. Yazdavar, H.S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, A. Sheth, Semi-supervised approach to monitoring clinical depressive symptoms in social media, in: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Network Analysis and

Mining, in: IEEE/ACM International Conference on Advances in Social Network Analysis and Mining, vol. 2017, 2017, pp. 1191–1198.

[34] K. Kroenke, R.L. Spitzer, J.B. Williams, B. Löwe, The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review, Gen. Hosp. Psychistry 32 (2010) 345–359.

[35] M. Kang, G. Choi, H. Jeon, J.H. An, D. Choi, J. Han, CURE: Context- and uncertainty-aware mental disorder detection, in: Y. Al-Onaizan, M. Bansal, Y.N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 17924–17940, http://dx.doi.org/10.18653/v1/2024.emnlp-main.994, https://aclanthology.org/2024.emnlp-main.994/.

[36] Z. Guo, N. Ding, M. Zhai, Z. Zhang, Z. Li, Leveraging domain knowledge to improve depression detection on chinese social media, IEEE Trans. Comput. Soc. Syst. 10 (2023) 1528–1536.

[37] S. Ghosh, A. Ekbal, P. Bhattacharyya, What does your bio say? inferring twitter users' depression status from multimodal profile information using deep learning, IEEE Trans. Comput. Soc. Syst. 9 (2022) 1484–1494.

[38] S. Ghosh, T. Anwar, Depression intensity estimation via social media: A deep learning approach, IEEE Trans. Comput. Soc. Syst. 8 (2021) 1465–1474.

[39] S.T. Sadasivuni, Y. Zhang, A new method for discovering daily depression from tweets to monitor peoples depression status, in: 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence, HCCAI, 2020, pp. 47–50.

[40] L. Belcastro, R. Cantini, F. Marozzo, D. Talia, P. Trunfio, Detecting mental disorder on social media: a chatgpt-augmented explainable approach, 2024, arXiv:2401.17477.

[41] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, S. Ananiadou, Towards interpretable mental health analysis with large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 6056–6077.

[42] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 328–339, http://dx.doi.org/10.18653/v1/P18-1031, https://aclanthology.org/P18-1031.

[43] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Fine-tuning a masked language model, 2022.

[44] M. Aragon, A.P. Lopez Monroy, L. Gonzalez, D.E. Losada, M. Montes, DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 15305–15318, http://dx.doi.org/10.18653/v1/2023.acl-long.853, https://aclanthology.org/2023.acl-long.853.

[45] B. Kim, H. Kim, G. Kim, Abstractive summarization of reddit posts with multi-level memory networks, 1 (2019) 9–1260, http://dx.doi.org/10.18653/v1/N, https://aclanthology.org/N19-1260.

[46] D. Losada, P. Gamallo, Evaluating and improving lexical resources for detecting signs of depression in text, Lang Resour. Eval. 54 (2020) 1–24, http://dx.doi.org/10.1007/s10579-018-9423-1.

[47] J. Parapar, P. Martín-Rodilla, D.E. Losada, F. Crestani, Overview of eRisk 2023: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2023, pp. 294–315.

[48] J. Parapar, P. Martín-Rodilla, D.E. Losada, F. Crestani, Erisk 2024 : Depression, anorexia, and eating disorder challenges, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 474–481.

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An Imperative Style, High-Performance Deep Learning Library, 32, Advances in Neural Information Processing Systems, 2019, pp. 8024–8035.

[50] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019, arXiv preprint arXiv:1901.11196.

[51] V. Kumar, A. Choudhary, E. Cho, Data augmentation using pre-trained transformer models, 2020, arXiv preprint arXiv:2003.02245.

[52] S. Kobayashi, Contextual augmentation: Data augmentation by words with paradigmatic relations, 2018, arXiv preprint arXiv:1805.06201.

[53] D.E. Losada, F. Crestani, J. Parapar, Overview of erisk: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th Interna- tional Conference of the CLEF Association,, CLEF 2018, Avignon, France, 34, 2018, pp. 3–361.

[54] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, M. Mitchell, Clpsych 2015 shared task: Depression and PTSD on twitter, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal To Clinical Reality, 2015, pp. 31–39.