Artificial Intelligence In Medicine 168 (2025) 103217

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Adapting language models for mental health analysis on social media

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, Santiago de Compostela, 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:
Social media
Mental health
Anorexia
Depression
Gambling
Self-harm
Language models
Adapters

ABSTRACT

In recent years, there has been a growing research interest focused on identifying traces of mental disorders through social media analysis. These disorders significantly impair millions of individuals' cognitive and behavioral functions worldwide. Our study aims to advance the understanding of four prevalent mental disorders: Anorexia, Depression, Gambling, and Self-harm. We present a comprehensive framework designed for the domain adaptation of models to analyze and identify signs of these conditions on social media posts. The language models' adapting strategy consisted of three key stages. First, we gathered and enriched substantial data on the four psychological disorders. Second, we adapted the different models to the language used to discuss mental health concerns on social media. Finally, we employed an adapter to fine-tune the models for multiple classification tasks (specific to each mental health condition). The intuitive idea is to adapt a language model smoothly to each domain. Our work includes a comparative study of different language models under in- and cross-domain conditions. This allows us to, for example, assess the ability of a depression-based language model to detect signs of disorders such as anorexia or self-harm. We show that the resulting mental health models perform well in early risk detection tasks. Additionally, we thoroughly analyze the linguistic qualities of these models by testing their predictive abilities using conventional clinical tools, such as specialized questionnaires. We rigorously examine the models across multiple predictive tasks to provide evidence of the adaptation approach's robustness and effectiveness. Our evaluation results are promising. They demonstrate that our framework enhances classification performance and competes favorably with state-of-the-art models.

1. Introduction

Mental disorders span a multifaceted array of conditions that profoundly impact cognitive, emotional, and behavioral well-being [1]. A wide variety of mental disorders, from mood-related conditions like depression or anxiety to more severe disorders such as schizophrenia, present a broad range of challenges and shape how individuals perceive themselves, interact with others, and interpret their surrounding world.

Individuals suffering from mental disorders may experience different symptoms, including persistent sadness, fear, intrusive thoughts, and impaired judgment [2]. Understanding the complexities of mental disorders is essential for promoting awareness, destigmatizing misconceptions, and fostering support for those enduring these problems. For instance, a recent study of Psychopathology conducted by the American Psychology Association indicates a rise in the need for mental health care services [3]. Factors such as economic pressures, the worldwide pandemic, population expansion, and climate change are all deemed to contribute to the escalation of mental health disorders. Hence, there is a growing demand for novel instruments to identify early symptoms of psychological concerns and track the evolution of mental disorders, helping in prevention before the condition worsens.

User-generated content is intensively published on social media, allowing researchers to explore how individuals confront different challenges. Many people utilize online platforms to share their daily experiences and essential events openly. Some individuals take advantage of the anonymity of these environments, intending to discuss mental health concerns candidly and seeking support [4,5]. Our objective in this study is to advance in the identification of indicators of mental disorders through automated analysis of social media posts. We focus on four prevalent mental disorders: Anorexia nervosa, major Depression disorder, self-injury (Self-harm) disorder, and Gambling disorder. The

* Corresponding author.
E-mail address: ezra.aragon@usc.es (M.E. Aragón).

https://doi.org/10.1016/j.artmed.2025.103217
Received 12 July 2024; Received in revised form 22 April 2025; Accepted 3 July 2025
Available online 16 July 2025

World Health Organization (WHO) has recognized that these four conditions hold significant importance and manifest prominently in contemporary society. These mental health challenges not only exhibit a noteworthy prevalence in our society but have also shown a concerning upward trend over the years.[1] Furthermore, a significant portion of individuals suffering from these conditions lack access to adequate care and support systems. This highlights a critical gap in mental health services that warrants urgent attention and intervention [6]. Through this research, we aim to contribute valuable insights into the multifaceted nature of these mental disorders. We are fostering a better understanding that can lead to more accessible and effective mental health care solutions.

In our study, we created and made publicly available robust language models focused on these critical mental health disorders.[2] We also compared them under in-domain and cross-domain early risk detection tasks, shedding light on the connection between different models and their predictive abilities. The cross-domain transferability of these language models has been scarcely studied in the literature. This is unfortunate, as mental health annotations do not abound. Thus, it is crucial to understand the viability of transferring models among psychological disorders. Furthermore, confronting these models with real clinical instruments, such as specialized questionnaires, is essential. This enables the examination of relevant language patterns. We contribute towards this goal by studying the models' predictions when prompted with relevant expressions derived from the medical questionnaires. This helps to understand the extent to which the language models integrate specialized knowledge. Overall, we expect to contribute to developing new technologies capable of alerting about the emergence of mental health issues and providing supportive evidence and explanations.

To build our models, we perform domain adaptation [7] and incorporate data collected and augmented from areas pertinent to this study (namely, Anorexia, Depression, Gambling, and Self-harm). Domain adaptation takes an already trained model and further refines it using a relatively modest corpus tailored to this specialized domain [8]. Our proposed framework adapts transformers to detect indicators of mental disorders following three key steps. Initially, we gather data relevant to each domain and employ a data augmentation technique to enrich the initial dataset. Next, we further teach the model using linguistic patterns from individuals exhibiting symptoms of mental disorders on social media platforms. Last, we integrate an adapter module [9] to facilitate the model's efficient adaptation to the task. This serves to smoothly fine-tune the models, taking advantage of the efficiency and adaptability of the adapter's technology, especially when datasets are small. We can summarize our contributions as follows:

- We introduce a simple yet powerful framework designed for the domain adaptation of models to identify markers of mental disorders within social media.
- We rigorously evaluate through empirical assessment the proposed approach, offering both quantitative metrics and qualitative insights to underscore the robustness of the approach in detecting signs associated with Anorexia, Depression, Gambling, and Self-harm.
- We evaluate the models' capacities to do in-domain and cross-domain predictions and delve into their relative merits and weaknesses when presented with evidence posted by users suffering from distinct mental health conditions. Furthermore, we also evaluate the models following a cross-platform approach, where we measure the transferability of the models from one social media platform to another.
- We gauge the models with textual data from real clinical questionnaires and assess the models' abilities to reproduce the specialized language utilized in such medical instruments.

The remainder of the paper is organized as follows. Section 2 provides an overview of the literature on detecting mental health disorders using social media data. Section 3 elaborates on the steps of our framework, offering technical details of our approach. Section 4 explains the experimental settings in detail, while Section 5 offers a comprehensive exposition of our experiments and results, shedding light on our main findings. Section 6 includes a qualitative analysis of the models to deepen the understanding of their performance and key features. Section 7 presents a set of cross-platform experiments, and Section 8 contains a general discussion of the evaluation results reported in the paper. Section 9 addresses the limitations and ethical considerations inherent to our work. Last, Section 10 draws our primary conclusions, summarizing the key insights gleaned from our study.

2. Related work

Recent studies have used social media platforms as a valuable resource for examining the manifestations of various mental disorders. Most of these studies employed automatic or semi-automatic techniques for data collection [10]. For example, many research projects focused on identifying users who openly disclosed a clinical diagnosis of a mental disorder in their public posts. Given this *positive group of users*, their entire threads of messages (or specific segments) are exploited for analysis [11]. Some authors developed methods for automatically monitoring individuals through information provided in their Twitter profile descriptions [12]. The subsequent analysis of their social interactions revealed intriguing patterns in tweeting preferences, language usage, mortality-related expressions, and emotional words. In [13], the authors investigated the role of personal statements in detecting depression through social media texts. Inspired by psychological studies about language use and individual focus, this paper hypothesized that user-generated texts with a strong self-focus may reveal signs of depression. This team thus proposed an approach emphasizing personal statements in textual representations. Their findings indicated that pronoun-rich phrases, especially from the *"I"* and *"you"* families, offer discriminative traces of depression.

Also, concerning language utilization, certain studies have utilized conventional classification algorithms and examined words and word sequences as distinctive features [14–16]. This class of analytical approaches is often oriented to comparing the prevalent word sets employed by individuals with mental disorders against those used by healthy users [17]. However, a drawback of this methodology is the commonly observed similarity between the vocabularies of both user groups [18,19]. Similarly, [20] presented a thorough examination of depression detection using machine learning on Twitter data and contributed significantly by building Arabic and English depression corpora, thus addressing the need for diverse language-specific datasets. This research study evaluated text pre-processing alternatives, feature extraction methods, and multiple classifiers through extensive experiments, offering insights into depression severity prediction.

Lately, there has been a growing interest in employing ensemble techniques that blend the aforementioned representations into diverse deep neural models [16,21] and in exploiting specialized questionnaires [22] or diagnostic and statistical manuals of mental disorders [23]. Related to this, Masood [24] proposed a methodology rooted in neural networks, multi-task learning, domain adaptation, and Markov models for early detection of signs of anorexia and self-harm.

Recent advances in natural language processing, particularly through transformer-based models like BERT and RoBERTa, have significantly improved the automatic detection and assessment of mental health conditions from user-generated text. Several studies have demonstrated the potential of these models in accurately identifying

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[2] Upon acceptance of the paper, we will release all models in the Hugging Face platform.

symptoms of mental disorders, emotional states, and behavioral risks. For example, Hossain et al. [25] designed Opinion-BERT to enhance standard BERT architecture by integrating opinion embeddings for multi-task learning, achieving high sentiment and mental health status classification performance. Similarly, Pourkeyvan et al. [26] exploited social media data and pre-trained BERT models, yielding effective mental disorder prediction, even with minimal input, such as user bios. Further improvements have been observed by integrating optimization strategies into BERT and RoBERTa [27], leading to enhanced data filtering and better classification accuracy. In clinical contexts, models trained on specialized corpora have shown strong performance in identifying nuanced linguistic indicators of mental illness, underscoring BERT's role as a valuable diagnostic aid [28]. Extending this work to multilingual healthcare data, Guardian-BERT [29] uses domain-adapted pre-training on Spanish electronic health records to detect non-suicidal self-injury and suicidal behavior. This approach also incorporates risk factor analysis to enhance model interpretability. These studies confirm the transformative potential of NLP-driven approaches for scalable, accurate, and explainable mental health screening.

Other studies have focused on training language models tailored to particular domains [30]. For instance, SciBERT [31] utilized an extensively annotated dataset comprising scientific information to fine-tune BERT specifically for scientific contexts. This adaptation demonstrated enhancements over the standard BERT model across various classification tasks. Similarly, BioBERT [32] underwent pre-training on vast biomedical corpora, surpassing the performance of the original BERT model in numerous biomedical text analysis tasks. In [33], the authors proposed a double-domain adaptation of a language model, adapting the model to social media language and then to the mental health domain. The experiments with the resulting model, DisorBERT, suggested that combining domain adaptation with lexical knowledge helps to detect traces of mental disorders. We recommend consulting comprehensive reviews on early risk detection [5] or surveys on computational methods for online mental state assessment [4] to explore this topic further.

Although many isolated studies have analyzed specific mental health conditions using certain computational tools, a complete picture of their effectiveness is still lacking. This paper extends and combines empirical evidence regarding language models for mental health analysis. We also provide a comprehensive quantitative and qualitative analysis of models tailored to four key psychological disorders. Our study includes in-domain and cross-domain evaluation conditions, and we also assess the models' abilities when prompted by expressions derived from established clinical resources. Furthermore, our methodology incorporates data collection and task adaptation through adapters, which has shown to be advantageous for fine-tuning, especially if datasets are not big enough [34]. We elaborate on this strategy in the following section. To the best of our knowledge, this is the first study that evaluates language models tailored to multiple psychological conditions under such a highly diversified set of evaluation conditions.

3. Proposed approach

This section describes our framework for building and adapting language models for different classes of mental health screening. The approach has three main stages: data collection, domain, and task adaptation. The first stage collects data related to mental health from multiple social media sources. The second stage adapts language models from a general domain (e.g., BERT or RoBERTa) to a more specialized mental health domain (using data collected from relevant sources). Next, the third stage leverages an adapter to specialize the models to a specific mental disorder detection task. Fig. 1 depicts the whole process, starting with data collection and passing through the domain and task adaptation stages.

We initiate the process by retrieving data from Reddit. The extracted publications are related to some of the four mental health disorders. Next, we enhance the dataset using back translation techniques. For domain adaptation, we adhere to the methodology suggested by [7,35], which extends the pre-training of BERT or RoBERTa by fine-tuning the masked language model for more epochs. This process uses the previously acquired and enhanced Reddit data and adapts the model to a mental health disorder. The goal is to refine the language model to capture the language used (and the dominant contexts) in Reddit's mental health discussions. Finally, to tackle the problem of detecting signs of mental disorders, we train the models using an adapter for each specific task (e.g., detecting early signs of anorexia). This training process is crucial for optimizing model performance in the downstream classification task. In this stage, we train the models during three epochs, with a batch size of 128 and a learning rate of $2e^{-5}$. We used an NVIDIA Tesla V100 32 GB SXM2 GPU for all training processes. The specifics of each step are explained below.

3.1. Data collection and augmentation

The first step consists of gathering data related to multiple mental disorders from social media. To that end, we explored Reddit communities tightly associated with various mental health conditions. More specifically, we selected several subreddits related to Anorexia, Depression, Gambling, and Self-harm,[3] and extracted multiple posted publications from each of these subreddits. It is important to note that the limitations set by Reddit constrain the data acquisition process. This social media platform restricts us to downloading a maximum of 1000 posts from each subreddit. To increase the number of extracted publications, we performed two downloading rounds spaced approximately one month apart. Each specific subreddit group contains highly focused communities and more general subreddits related to broader mental health topics.

We employed a data augmentation method to increase the size of the datasets further by adding additional textual examples. Data augmentation offers a suite of techniques to artificially expand datasets by generating additional data points from existing ones. We specifically utilized back translation [36]. Back translation is the process of translating original content from the source language to a specific target language and, next, translating it back to the source language. For instance, if a piece of content is translated from English to Spanish, the translator would produce a back translation in English. This approach creates additional representations for the original sentences. Such an approach promotes clarity regarding the intended meaning of the source text. Back translation proves beneficial, mainly when dealing with content such as slogans, titles, product names, and puns, where the implied meaning in one language might not resonate effectively in another. This technique is helpful in social media, where users often employ colloquial language. In this phase, we performed a multi-step process involving the translation of collected posts into Spanish,[4] followed by a translation back to English.[5] The upper part of Fig. 1 shows an example of this procedure, where the post *"I've been struggling a lot lately"* is augmented to *"I've been fighting a lot lately"*, which maintains the original meaning but employs slightly different wording. Table 1 reports each disorder's counts of original and new posts.

3.2. Domain adaptation

This step involves adapting the general pre-trained language models (base BERT and RoBERTa) to accommodate the language commonly used within the mental health domain, particularly in social media scenarios. We employed the expanded dataset previously created (data

[3] In Appendix we detail the group of subreddits chosen for each disorder.
[4] model used: https://huggingface.co/Helsinki-NLP/opus-mt-en-es.
[5] model used: https://huggingface.co/Helsinki-NLP/opus-mt-es-en.

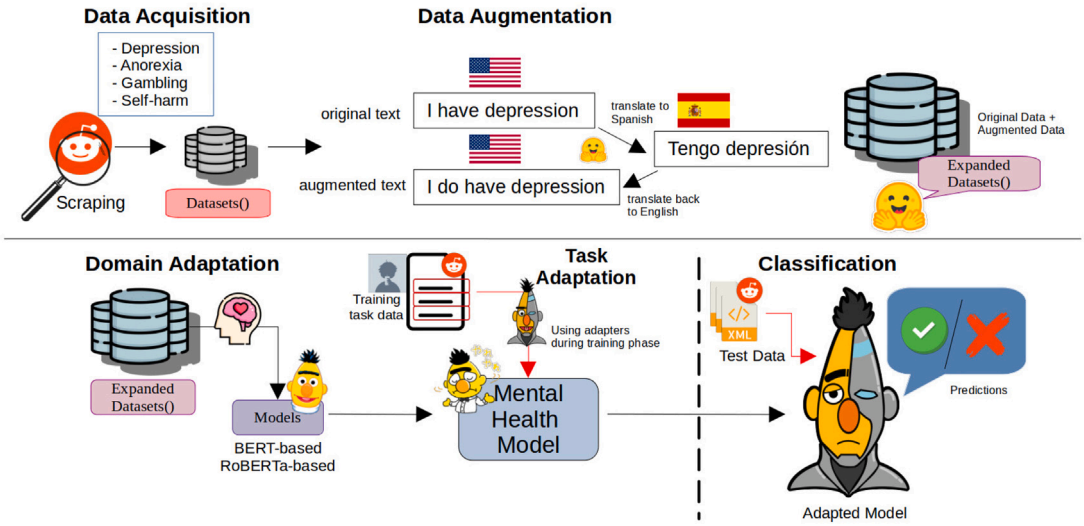

Fig. 1. General diagram of the adaptation process. It starts with data collection and augmentation. Then, a base language model is learned from Reddit (using multiple mental health subcommunities). Last, the language model training incorporates an adapter for efficient task adaptation.

Table 1
Number of posts and size (MBs) of the original dataset collected from Reddit and the new examples included in the augmentation process.

Disorder	Original	New
Anorexia	24,831/32,2 MB	24,510/13,3 MB
Depression	48,261/44,9 MB	40,170/21,5 MB
Gambling	22,830/13,1 MB	22,820/8,2 MB
Self-harm	22,057/19,5 MB	22,050/11,5 MB

collected and augmented) to accomplish this adaptation. The adaptation consists of continuing the training of the base models by refining them for additional epochs [7]. In our experiments, we built one model for each mental health disorder and a generalist model (where domain adaptation was performed using all available data). This led to the thorough study of five different language models, with (i) in-domain and cross-domain early risk detection experiments and (ii) other disorder-related language prediction tasks.

To construct the models, we aggregated all posts for each set of subreddits and partitioned the dataset into chunks of equal size, each containing 128 examples. The training procedure of the language model utilized language model masking and next-sentence prediction, as traditionally employed for transformer architectures [37].

This methodological design aims to provide valuable insights into the nuances of language and the intricacies of language use and mental health disorders. For example, a thorough analysis of these five language models reveals how individuals suffering from different psychological disorders express themselves and identifies specific language trends across multiple domains and subcommunities. It is worth noting that the datasets contain a wide range of publications posted by diverse social media users. For instance, some users may express negative emotions in their posts, but they may not necessarily be experiencing psychological distress.

3.3. Task adaptation

Specific language learning tasks encounter difficulties due to the scarcity of large annotated datasets. This presents a significant obstacle in training models like deep neural networks. With few instances for learning, the model's capacity to grasp the underlying data distribution becomes more challenging, resulting in biased or inaccurate predictions. To address this limitation, we opted for adapters, a lightweight neural network component integrated into the existing pre-trained model structure [9]. Adapters serve to streamline the fine-tuning process of the pre-trained model for specific downstream tasks, minimizing extensive alterations of the original parameters. Their primary advantage lies in their efficiency and adaptability, and they have proven effective when the final adaptation is in tasks with small data, such as clinical ones. Rather than fine-tuning the entire pre-trained model, which can be resource-intensive and may erase previously acquired knowledge, adapters offer a modular and flexible solution. They enable task-specific adjustments without significantly disrupting the established representations of the original model. Besides, adapters can be trained with fewer data points than a complete model. Thereby reducing training costs and enhancing the model's scalability. This allows us to save resources by using the same base model for different tasks and only changing the adapter for specific tasks [34].

Note that adapters usually entail a limited set of extra parameters in contrast to the entire model, enhancing computational efficiency. They are integrated into particular layers of the pre-trained model, adapting representations to suit the demands of downstream tasks. This facilitates swift adjustment to novel tasks and domains while retaining the knowledge and skills gained during pre-training. In our experiments, we employed a parallel adapter [38] and incorporated it into the five previously discussed models. At the end of this stage, the framework's output consists of a fine-tuned model that can be exploited for multiple prediction tasks. In particular, we use it for classification.

4. Experimental settings

4.1. Datasets

For the evaluation, we utilized the datasets derived from the eRisk 2019–2023 evaluation tasks [39–43]. These evaluation campaigns were structured to foster the identification of early indicators related to anorexia, depression, gambling, and self-harm. In Table 2, we present an overview of the datasets, including details about their composition and the distribution of classes.

These eRisk datasets contain the histories of publications (posts or comments) from numerous Reddit users. Each collection has two classes: (i) positive users (affected with conditions such as anorexia, depression, gambling, or self-harm), and (ii) a control group containing individuals not afflicted by these mental disorders. The creators of these collections identified positive users through searches for explicit mentions of a diagnosis (i.e., a user openly declaring that a medical specialist had diagnosed them). Ambiguous expressions like *"I think I*

Table 2
Datasets used for experimentation. P refers to the positive users, and C refers to the control users.

	Train		Test	
	P	C	P	C
Anorexia 2019				
# users	61	411	73	742
Avg # pubs	407.8	556.9	241.4	745.1
Avg # words/pub	37.3	20.9	37.2	21.7
Depression 2022				
# users	214	1493	98	1302
Avg # pubs	440.9	660.8	360.5	527.8
Avg # words/pub	27.5	22.75	27.4	23.5
Gambling 2023				
# users	245	4182	103	2071
Avg # pubs	256.9	499.63	327.3	516.2
Avg # words/pub	30.5	21.1	28.9	20.4
Self-harm 2021				
# users	145	618	152	1296
Avg # pubs	128.4	412.0	336.2	531.5
Avg # words/pub	22.4	15.2	26.03	20.74

have anorexia/depression" or *"I am anorexic/depressed"* were not considered as an indication of a diagnosis. The control group consists of randomly selected users from multiple subReddits, as well as some individuals who frequently engage in discussions related to anorexia, depression, gambling, or self-harm. For instance, the control group includes some expert clinicians actively participating in mental health subreddits (e.g., providing support and advice to others). Consequently, risk detection technology cannot rely solely on discerning the topic of conversations; instead, a more subtle understanding of user interactions and preoccupations is needed. This data compilation approach gives a heightened level of realism to these datasets.

4.2. Training configurations

Pre-processing: We conducted basic text pre-processing by converting all words to lowercase and eliminating certain elements such as URLs, emoticons, and hashtags. This initial step removes non-essential parts that may not contribute to the analysis.

Training and predictions: Let $U = \{u_1, u_2, \ldots, u_M\}$ denote the set of users, where each user u_i is associated with a sequence of publications $P_i = \{p_1, p_2, \ldots, p_{L_i}\}$, and p_j is the jth publication of user u_i. Each user is labeled either as positive (e.g., at risk) or control, with corresponding binary labels $y_i \in \{0, 1\}$. Each publication p_j is further processed into a fixed-length segment by truncating or padding its tokenized representation to a length of $N = 35$. This ensures uniformity across all input texts while preserving the individual nature of each publication. During training, each such segment inherits the label of its originating user, resulting in a training set composed of segment–label pairs (s_{ij}, y_i), where s_{ij} is the tokenized and length-normalized version of the jth publication by user u_i. Each segment is encoded using the BERT/RoBERTa models to obtain the vector representation, and a binary classifier is trained on these vectors to distinguish between positive and control users. During prediction, each test user u_t is processed similarly to yield a set of segments $S_t = \{s_1, \ldots, s_{L_t}\}$, and the classifier is applied to each segment to produce individual predictions $\hat{y}_{tj} = f(s_{tj})$. The final user-level prediction $\hat{y}_t$ is determined via a majority voting strategy: a user is labeled as positive if more than half of the segments are classified as positive (i.e., positive if $\frac{1}{L_t}\sum_{j=1}^{L_t} \hat{y}_{tj} > 0.5$, and negative otherwise). This aggregation method assumes that at-risk users tend to consistently express worrying signals across multiple publications, enabling reliable identification at the user level.

Parameters: We employed the frameworks offered by HuggingFace v4.24.0 [35] and PyTorch v1.13.0 [44] for model development. Specifically, during the training phase, we adopted a batch size 256, leveraged the Adam optimizer with a learning rate set to $1e^{-5}$, and utilized cross-entropy as the loss function. The training process spanned three epochs and was conducted on a GPU NVIDIA Tesla V100 32 GB SXM2 to expedite computation and model convergence.

4.3. Baseline approaches

BERT: This baseline method implements a classification model based on BERT,[6] a state-of-the-art language representation model. We fine-tuned a BERT classification model using each training set. This fine-tuning enhances the general model and adapts it to our specific tasks.

RoBERTa: This is a Robustly Optimized BERT Approach.[7] It is a variant of the BERT model. RoBERTa builds upon BERT's architecture and pre-training techniques but incorporates several modifications and optimizations to improve its performance. Similar to BERT, we adapted this model to each task through fine-tuning.

MentalBERT: This is a language model pre-trained for the mental healthcare domain. It was built from many sentences extracted from Reddit [45]. Similar to previous models, we fine-tuned this model over each training set.

MentalRoBERTa: A variation of MentalBERT using a RoBERTa model trained with mental health-related posts collected from Reddit.

4.4. Adapted models

The names of the models refer to the specific corpus (see Appendix) used for the domain adaptation step. For example, AnorBERT was built with the anorexia data, while DepBERT was constructed with the depression data. The complete list of models is as follows:

AnorBERT: base model (BERT) + domain adaptation using the anorexia data from Reddit.

AnorRoBERTa: base model (RoBERTa) + domain adaptation using the anorexia data from Reddit.

DepBERT: base model (BERT) + domain adaptation using the depression data from Reddit.

DepRoBERTa: base model (RoBERTa) + domain adaptation using the depression data from Reddit.

GambBERT: base model (BERT) + domain adaptation using the gambling data from Reddit.

GambRoBERTa: base model (RoBERTa) + domain adaptation using the gambling data from Reddit.

SHBERT: base model (BERT) + domain adaptation using the self-harm data from Reddit.

SHRoBERTa: base model (RoBERTa) + domain adaptation using the self-harm data from Reddit.

WholeBERT: base model (BERT) + domain adaptation using all Reddit datasets.[8]

5. Evaluation results

Table 3 reports the results of the models without using adapters (we defer the comparison of models with and without adapters to the end of this section). The table presents results for the four classification tasks, comparing different variants and baseline methods. To put these results into context, we also report (third row) the average results of the research teams participating in the eRisk campaigns. However,

[6] https://huggingface.co/google-bert/bert-base-uncased.
[7] https://huggingface.co/FacebookAI/roberta-base.
[8] Due to the less effective results obtained with the RoBERTa-based models, we decided to test only the BERT model with the entire dataset.

Table 3
In-domain experiments (without adapters). F1, Precision (P), and Recall (R) results over the positive class in four eRisk tasks.

	Anorexia			Depression			Gambling			Self-Harm		
	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Baselines	In-domain											
BERT	.759 ± .007	.692 ± .040	.845 ± .044	.476 ± .006	.378 ± .007	.643 ± .000	.745 ± .019	.739 ± .043	.752 ± .007	.552 ± .022	.398 ± .027	**.899 ± 0.020**
RoBERTa	.721 ± .020	.602 ± .031	**.899 ± .008**	.445 ± .020	.352 ± .033	.636 ± .015	.647 ± .005	.559 ± .017	.767 ± .019	.529 ± .031	**.706 ± .057**	.428 ± .057
MentalBERT	.769 ± .028	.806 ± .127	.758 ± .124	.537 ± .004	.570 ± .056	.514 ± .048	.709 ± .068	.743 ± .017	.793 ± .100	.633 ± .004	.647 ± .011	.618 ± .000
MentalRoBERTa	.778 ± .022	.766 ± .141	.822 ± .131	.508 ± .008	.441 ± .012	.599 ± .015	.723 ± .007	.763 ± .037	.689 ± .035	.448 ± .010	.722 ± .015	.325 ± .008
eRisk avg	.417 ± .208	.412 ± .222	.588 ± .284	.324 ± .145	.226 ± .153	**.838 ± .166**	.391 ± 0.351	.416 ± .409	**.848 ± .228**	.359 ± .146	.278 ± .170	.774 ± .207
Our models												
AnorBERT	.777 ± .006	.773 ± .070	.790 ± .068	.494 ± .014	.399 ± .018	.650 ± .012	.743 ± .006	.733 ± .005	.754 ± .006	.631 ± .041	.524 ± .076	.809 ± .063
AnorRoBERTa	**.790 ± .014**	**.813 ± .013**	.768 ± .037	.455 ± .010	.363 ± .016	.639 ± .016	.646 ± .005	.574 ± .017	.767 ± .020	.535 ± .021	.655 ± .054	.454 ± .024
DepBERT	.766 ± .005	.725 ± .032	.813 ± .034	.543 ± .005	.525 ± .021	.565 ± .035	.734 ± .004	.722 ± .010	.748 ± .017	.657 ± .008	.591 ± .015	.741 ± .027
DepRoBERTa	.738 ± .013	.633 ± .025	.886 ± .016	.462 ± .013	.364 ± .020	.633 ± .017	.697 ± .010	.680 ± .016	.715 ± .011	.517 ± .027	.653 ± .044	.432 ± .050
GambBERT	.773 ± .006	.741 ± .011	.808 ± .00	.488 ± .020	.391 ± .022	.650 ± .012	**.768 ± .005**	**.806 ± .025**	.728 ± .024	.658 ± .006	.576 ± .010	.768 ± .020
GambRoBERTa	.740 ± .011	.636 ± .016	.886 ± .008	.462 ± .018	.369 ± .026	.619 ± .012	.652 ± .004	.571 ± .009	.761 ± .005	.531 ± .010	.677 ± .012	.436 ± .017
SHBERT	.771 ± .032	.721 ± .062	.831 ± .021	.493 ± .020	.399 ± .028	.646 ± .006	.739 ± .004	.727 ± .011	.751 ± .011	**.672 ± .007**	.615 ± .018	.741 ± .036
SHRoBERTa	.725 ± .015	.609 ± .024	.895 ± .008	.461 ± .016	.361 ± .024	.639 ± .016	.689 ± .017	.673 ± .013	.705 ± .024	.550 ± .008	.671 ± .032	.467 ± .017
WholeBERT	.776 ± .010	.742 ± .022	.813 ± .008	**.557 ± .004**	**.531 ± .013**	.585 ± .012	.751 ± .012	.753 ± .047	.751 ± .020	.667 ± .015	.655 ± .003	.680 ± .027

note that these teams often leveraged multiple sophisticated methods – such as ensemble techniques, annotating additional training data, or designing advanced neural architectures – and our intention is not to beat these models but to understand the relative merits of the focused language models.

This first table of results informs us about the **in-domain experiments**, where models were built using the training split from the same task. We ran each experiment thrice and reported average performances and standard deviations.

As a first observation, we can emphasize the improved performance of the adapted models compared to their base counterparts. This outcome underscores the significance of domain-specific models. Most of our proposed models outperform the baseline models regarding the F_1 score. The baseline models, which do not incorporate domain adaptation, tend to yield high recall but low precision. Among the baselines, MentalBERT and MentalRoBERTa performed better than the other models. Although the results of MentalBERT and MentalRoBERTa are close to those achieved by the adapted models, the adaptation seems crucial, as one of the adapted variants generally achieves the top performance. Overall, our models demonstrate an ability to adapt to the target type of language, identifying the corresponding signs of risk. For example, the GambBERT model yielded the highest F1 in the gambling detection task, and the SHBERT model got the highest F1 in the self-harm detection task. The best result for anorexia was obtained for the AnorRoBERTa model. For depression, the best model was the one trained with the whole data (but DepBERT's effectiveness was similar to that achieved by WholeBERT). This confirms that the corpora selected for the adaptations accurately capture language expressions relevant to these psychological disorders. Furthermore, our models' in-domain performance shows a good balance between precision and recall.

The in-domain results suggest a robust risk detection behavior, effectively identifying multiple indicators of psychological risks. These results are significant, especially in clinical screening tools, where high recall is crucial. However, it is essential to acknowledge potential scenarios where emphasizing precision might be preferable, such as in a social network targeting the riskiest behaviors. In such cases, adapting the models to prioritize precision may become necessary.

The effectiveness of these models underscores their robustness across diverse data domains, highlighting their potential for broader applicability in real-world scenarios. This indicates their predictive prowess and suggests a level of adaptability and generalization that sets them apart from the other baselines.

This comparison provides valuable insights into the effectiveness of different model configurations across multiple mental health screening tasks. It lays the groundwork for informed decision-making in the clinical domain (e.g., to guide preemptive measures). Expanding on our previous analysis, Fig. 2 presents a visualization of precision and recall, comparing our adapted models with the baselines. Our models exhibit a clustering tendency along the main diagonal, indicating a good balance between precision and recall. Conversely, the baselines often demonstrate a trade-off between these metrics, excelling in one dimension while lagging in another.

Table 4
Results including adapters. F1, Precision (P), and Recall (R) results over the positive class in four eRisk tasks.

Model/Metrics	F1	P	R
	Anorexia		
AnorBERT	.777 ± .006	.773 ± .070	**.790 ± .068**
AnorBERT+adapter	**.791 ± .007**	**.912 ± .030**	.699 ± .027
	Depression		
DepBERT	.543 ± .005	.525 ± .021	.565 ± .035
DepBERT+adapter	**.571 ± .014**	**.541 ± .013**	**.605 ± .021**
	Gambling		
GambBERT	**.768 ± .005**	**.806 ± .025**	**.728 ± .024**
GambBERT+adapter	.760 ± .005	.802 ± .035	.725 ± .034
	Self-harm		
SHBERT	**.672 ± .007**	.615 ± .018	**.741 ± .036**
SHBERT+adapter	.656 ± .007	**.644 ± .030**	.671 ± .039

An interesting pattern emerges when examining the performance in specific datasets. In Depression, Gambling, and Self-harm, the WholeBERT and BERT (domain-specialized) models display a notable cohesion, aligning closely with each other. This suggests that these models have similar predictive capabilities, with a shared effectiveness in capturing the intricacies of these disorders. This outcome possibly owes to the adaptability of these models to nuanced linguistic patterns and contextual cues of the Depression, Gambling, and Self-Harm datasets.

Let us now evaluate the **effect of adapters**. To that end, the comparison between the four focused models and their corresponding variants, which had adapters integrated at the task adaptation stage, is presented in Table 4. The integration of adapters results in a noticeable improvement in F1 performance for the anorexia and depression tasks. The adapter-based variants did not attain the highest F1 scores for self-harm and gambling, yet their performance was closely aligned with that of the non-adapter version. This confirms our hypothesis about the effectiveness of adapters, and these experiments represent the first application of adapters for datasets of this nature. In the following subsection, we further discuss the computational advantages of using adapters.

The results of **cross-domain experiments** are reported in Table 5. Here, we trained the adapted models using the labeled data from one task and evaluated the resulting classifiers against test examples from a different task. This helps to understand, for example, the transferability of a depression-based classifier for detecting signs of concern in individuals suffering from anorexia. Training models with examples from another task achieve lower performance than in-domain classifiers.

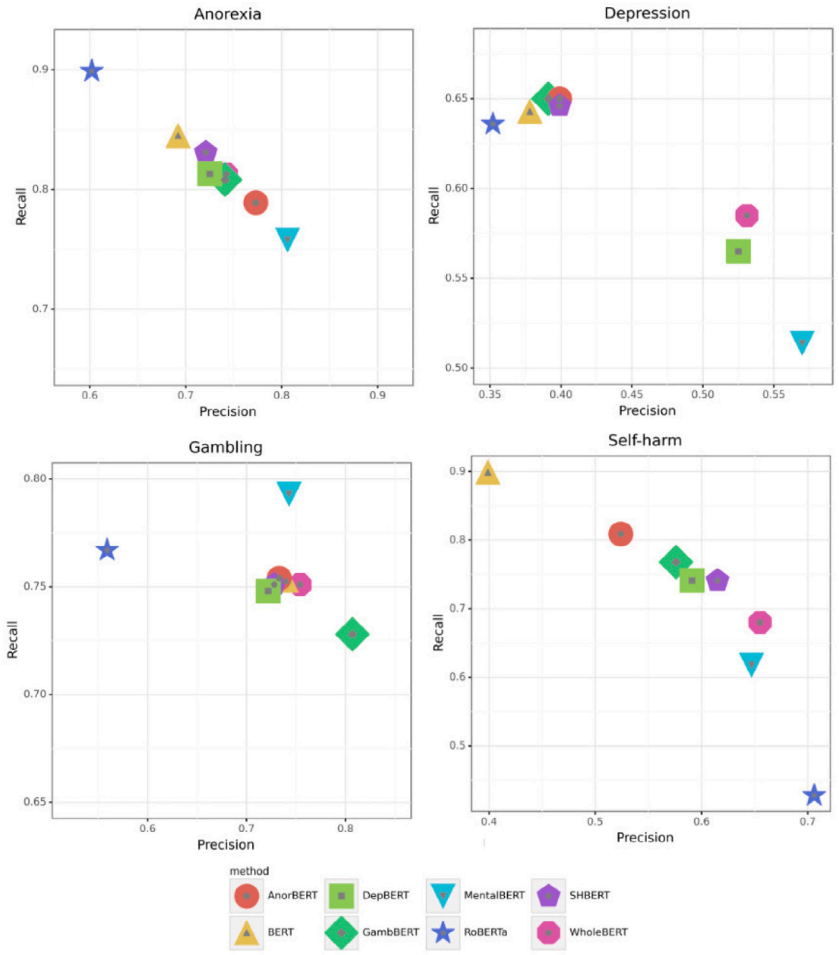

Fig. 2. Precision and Recall of the models for each task. We can observe that the adapted models tend to align in the main diagonal, showing a balance between the two metrics.

Table 5
Cross-domain F1 results. The models (rows) were trained with labeled examples from a different mental disorder (columns report the source of training data). For each column, we underline its best result. We mark in **bold** the overall best result for each task.

Task	Anorexia			Depression			Gambling			Self-Harm		
Model/Train data	Dep	Gamb	SH	Anor	Gamb	SH	Anor	Dep	SH	Anor	Dep	Gamb
AnorBERT	.595 ± .010	.321 ± .009	**.738 ± .002**	.503 ± .005	.404 ± .006	.496 ± .020	.139 ± .015	.114 ± .008	.221 ± .004	.582 ± .007	.511 ± .004	.116 ± .013
AnorRoBERTa	.372 ± .013	.274 ± .031	.669 ± .012	.428 ± .009	.396 ± .014	.453 ± .037	.135 ± .008	.107 ± .005	.098 ± .036	.557 ± .009	.307 ± .016	.104 ± .005
DepBERT	.575 ± .023	.314 ± .010	.707 ± .032	.476 ± .019	.412 ± .019	.478 ± .013	.143 ± .010	.104 ± .004	.230 ± .012	.587 ± .017	.573 ± .002	.119 ± .012
DepRoBERTa	.365 ± .011	.250 ± .011	.629 ± .018	.422 ± .007	.396 ± .016	.427 ± .014	.132 ± .019	.101 ± .007	.074 ± .029	.544 ± .005	.319 ± .009	.111 ± .012
GambBERT	.597 ± .018	.315 ± .023	.727 ± .004	.498 ± .002	.409 ± .005	.463 ± .019	.163 ± .022	.128 ± .011	**.230 ± .009**	.583 ± .003	.495 ± .008	.109 ± .001
GambRoBERTa	.374 ± .021	.262 ± .021	.650 ± .040	.434 ± .009	.396 ± .014	.472 ± .029	.167 ± .011	.118 ± .004	.121 ± .007	.558 ± .011	.325 ± .018	.111 ± .006
SHBERT	.567 ± .028	.308 ± .021	.728 ± .017	**.505 ± .001**	.413 ± .002	.479 ± .006	.155 ± .014	.109 ± .004	.228 ± .023	**.590 ± .001**	.500 ± .008	.112 ± .007
SHRoBERTa	.370 ± .010	.240 ± .036	.666 ± .024	.425 ± .016	.383 ± .006	.450 ± .021	.144 ± .006	.105 ± .007	.077 ± .043	.550 ± .013	.521 ± .026	.103 ± .006

This is a natural consequence of lacking textual examples of the same psychological disorder. Still, in some cases, the cross-domain performance is competitive, thus showing the potential for transferability. For example, the anorexia-labeled examples are effective for tuning the models to predict depression or self-harm. In contrast, the self-harm examples are suitable for tuning the models to predict anorexia. This could suggest that the language of people suffering from anorexia shares common psychological concerns and risks with people suffering from self-harm. In contrast, we can observe that all cross-domain models perform poorly on the gambling task. This suggests that the language of these individuals does not share many elements with those of the other disorders.

Note also that the specialized models (e.g., AnorBERT in Anorexia or GambBERT in Gambling) tend to obtain the best (or close to the best) result, even when fed with training data from another task. Again, this shows that the language learned in the adaptation phase helps the models to focus on relevant textual cues. Finally, we can observe that BERT models, in most cases, perform better than their RoBERTa counterparts.

5.1. Computational advantages of adapters

Our experiments show the effectiveness of adapters across different mental health screening tasks. Notably, one of their key advantages lies in their ability to optimize the utilization of pre-trained models. Adapters facilitate the integration of task-specific functionalities without altering the original architecture. This feature streamlines the adaptation process and enhances the model's versatility.

Moreover, adapters enable selective modification of specific layers. Unlike the conventional approach of fine-tuning the entire model, adapters offer a more nuanced strategy. They bolster the model's adaptiveness to novel tasks by enabling targeted adjustments while

Table 6
Percentage of training time required for the adapters (compared with the non-adapter setting).

Anorexia	Depression	Gambling	Self-harm
≈30–40%	≈11–18%	≈50–56%	≈45–52%

minimizing the computational overhead. In our experiments, we **only needed to train the 6.56% of parameters for BERT and 5.80% for RoBERTa** models. This translates to significantly reduced costs and memory requirements for training and storing the models, a crucial consideration in resource-constrained environments. Table 6 reports the percentage of training time required (compared to training the complete model) achieved by the adapter-based variants. We can observe a considerable reduction in training times, especially in the depression and anorexia tasks.

Adapters thus offer a compelling advantage regarding cost-effectiveness and allow the use of the same model in different tasks by switching the adapter only. Furthermore, they can be trained with less data than the entire model. To assess this advantage, we conducted an experiment where models were trained using varying percentages of training data ranging from 10% to 100%. The results, illustrated in Fig. 3, show that the focused models (with or without adapters) consistently outperform the BERT base model. This becomes particularly apparent in domains like anorexia, where we have less data for training. The adapter-based model achieves peak performance with training data ranging from 20% to 50%. These trends underscore the viability of adapters in scenarios where labeled data is limited. By demonstrating good performance with fewer training examples, adapters present an opportunity to address problems in domains where data availability is restricted. This opens exciting avenues for more efficient and cost-effective model training.

5.2. Early detection task

The eRisk evaluation campaign emphasized early risk detection and evaluated classification performance based on partial releases of the user's publication history. Under the early risk detection setting, the predictive algorithms could not access the complete sequence of users' publications. Instead, the posts published by test users were given in time-spaced rounds. We conducted a supplementary experiment to simulate this early detection scenario. The main goal was to understand the effectiveness of the classifiers when presented with partial representations of the test users' data. More specifically, we partitioned the test user data into distinct segments or "chunks" and measured the effectiveness of the predictions for each percentage of user data seen by the model. This helps to ascertain the robustness of our models in delivering early predictions. For example, a highly effective early detection algorithm would discern concerning signs of risk even when exposed to only 10% of the user's publications. This analysis thus provides insights into the temporal dynamics of our predictive models. It simulates their performance in real-time scenarios, where evidence comes naturally as time-spaced social media publications.

Each data chunk contains a sequence of the user's posts,[9] and we performed classification using the accumulated user-posted evidence at that point (using all posts from the current chunk and the earlier chunks). Consequently, with each additional chunk, the classifier benefits from accumulating evidence, leading to increasingly informed decisions. The classifier considers a user positive if more than 50% of the user's posts are categorized under the risk class. Fig. 4 plots the performance of our models with an increasing percentage of user posts (i.e., a growing number of chunks). Our proposed models perform

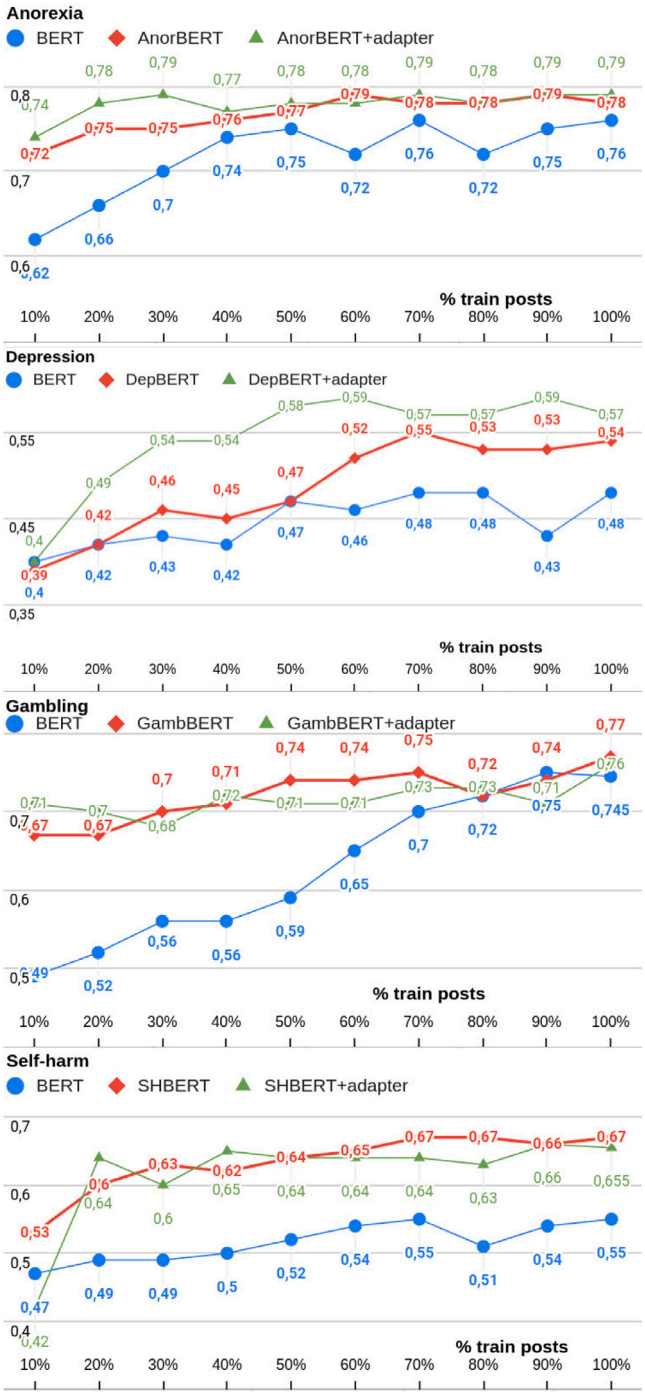

Fig. 3. F1 test performance with increasing percentages of training data. We can observe the effectiveness when using small data samples.

well at the early stages, notably for anorexia. This suggests that people suffering from anorexia quickly expose their symptoms. The adapter-based variants (green lines) are quite effective at detecting symptoms early. For instance, in the case of depression, the adapter-based model attained an F1 score of 0.49 using only the first available chunk (which contains only 10% of user-posted content). This model surpasses the second-best approach, which achieves an F1 score of 0.45 and is significantly better than the base model (F1 score of 0.37). Thus, the adapter models demonstrate competitive performance even with access to only a portion of the users' posts. This suggests that they can extract valuable insights from a limited subset of user posts.

[9] And the chunks are chronologically ordered following the original dates where the posts were published.

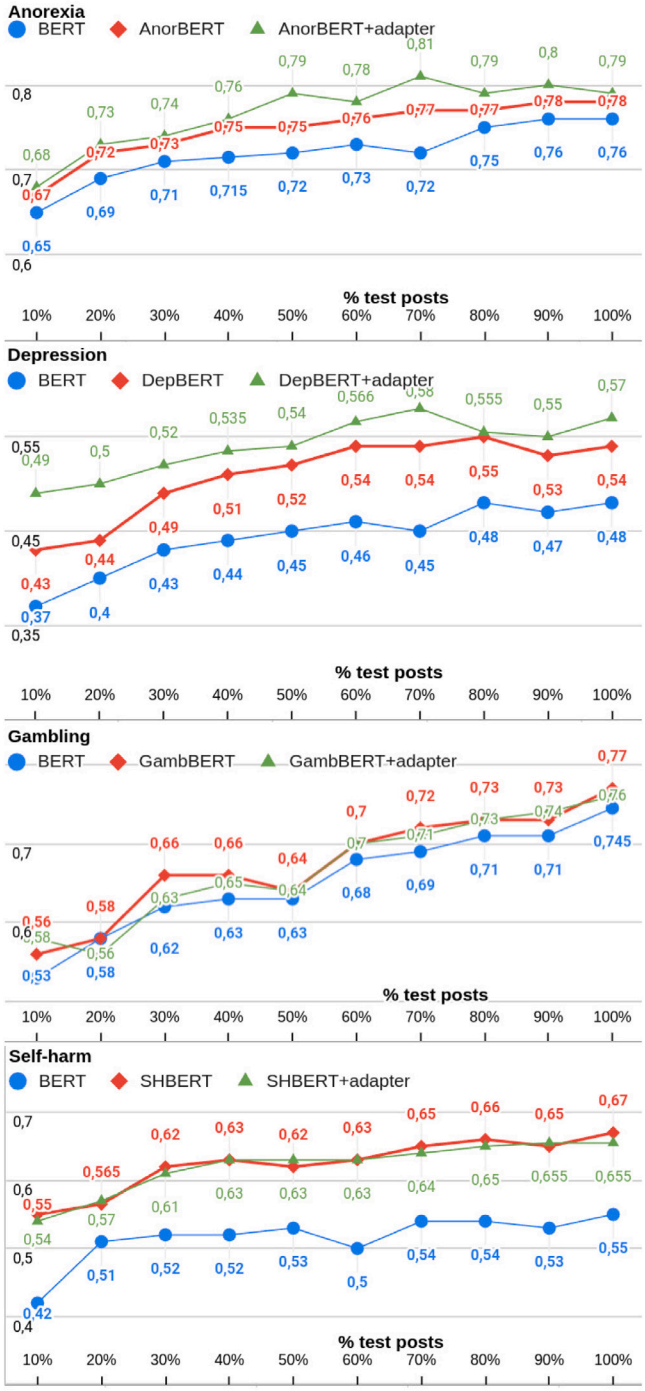

Fig. 4. F1 performance using different percentages (chunks) of test user's posted content. This aids in determining the reliability of our models in providing early predictions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Analysis of the models

The BERT base model was trained on a broad corpus to comprehend language at a global level. In contrast, our models are guided to the domain of mental health. In this section, we demonstrate the behavior of the trained models and highlight the types of textual expressions they tend to prioritize.

In this analysis, we assessed the word-generation capabilities of the models when presented with sentences containing masked words. To that end, it is essential to leverage standardized clinical instruments that contain cue words and expressions indicative of the corresponding psychological symptoms.

We utilized four questionnaires that aim to identify and measure the severity of typical symptoms of anorexia, depression, self-harm, and gambling, respectively. These self-report clinical instruments contain key textual expressions (e.g., about negative cognitions or feelings); for each item, we selected a core word, masked it, prompted the models with the masked item, and examined the word choices provided by each model. We used the Beck's Depression Inventory (BDI) for depression [46], the Eating Disorder Examination Questionnaire (EDE-Q) for anorexia [47], the Diagnostic Screen for Gambling Disorders (NODS-CLiP) for gambling [48], and the self-assessment self-harm test.[10] These questionnaires are professional psychological assessment tools designed to screen for potential disorders or addictions. They are concise questionnaires that help identify behaviors indicative of these problems. The tests focus on critical aspects of behavior to evaluate the risk of the disorder, aiming to prompt early intervention and support.

To quantify the effectiveness of the models in predicting this specialized language, we employed Mean Reciprocal Rank (MRR), a standard metric for evaluating the effectiveness of search systems in ranking correct answers. Essentially, MRR quantifies the ability of the models to find the correct answer among the candidates. Reciprocal Rank is calculated as the inverse of the rank of the correct answer, with the first rank receiving a value of 1, the second 1/2, the third 1/3, and so forth. The mean reciprocal rank is calculated by averaging the inverse of the ranks of the results for a sample of queries Q:

$$MMR = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (1)$$

where $rank_i$ refers to the rank position of the first relevant word for the ith query. We computed the MRR scores by averaging the reciprocal ranks obtained by each model across multiple masked sentences. Only the top 5 word predictions were considered (i.e., correct answers ranked at lower positions were assigned a reciprocal rank score equal to 0).

Table 7 presents the MMR results of the models for the four reference questionnaires. We observe that each disorder-focused model achieves the highest or the second-highest performance in predicting relevant words from its reference questionnaire. For example, AnorBERT was the best model in predicting relevant words from the Eating Disorder Examination Questionnaire. We can also highlight two other salient outcomes here. First, the base model yielded the poorest performance. This was expected since it was trained in a general domain. Second, the model trained with examples from all disorders (WholeBERT) attained the highest average performance. This underscores its capacity to screen signs of multiple mental disorders. Still, if the goal is to focus on a single mental health condition, then the specialized models often appear to be a better choice. Another interesting finding is that the depression-based model (DepBERT) was solid at predicting words from the BDI, but the anorexia-based model (AnorBERT) was even better. This underlines the oft-quoted prevalence of depression among individuals suffering from anorexia.

Nevertheless, the models struggled with several items from the four screening tools and sometimes obtained modest results, showing the task's difficulty. This is especially true in tasks such as gambling or self-harm, where the vocabulary is very specific, and models struggle to nominate the correct words.

To complement this analysis, Fig. 5 shows examples of sentences, masked words, and the models' predictions (sorted by decreasing likelihood). Our models estimate the correct answer more accurately than the base model. Notably, the selected answers often align with core

[10] https://www.mind.help/assessments/self-harm-test/.

Questionnaire	Example	BERT	AnorBERT	DepBERT	GambBERT	SHBERT	WholeBERT
EDE	Have you been deliberately trying to limit the amount of food you eat to influence your diet "Have you been deliberately trying to limit the amount of food you eat to influence your [MASK]"	? , ; ... behavior	**diet** mood health behavior behaviour	mood **diet** health behavior behaviour	behavior decisions health mood decisions	behaviour mood health **diet** behaviour	**diet** mood behavior health eating
BDI	I have lost all of my interest in other people. "I have lost all of my [MASK] in other people."	faith trust **interest** belief confidence	faith **interest** trust confidence belief	**interest** faith trust confidence hope	faith **interest** trust confidence belief	faith **interest** trust confidence belief	**interest** faith trust confidence hope
NODS-CLiP	Have you ever lied to family members, friends, or others about how much you gamble or how much money you lost on gambling? "Have you ever lied to family members, friends, or others about how much you [MASK] or how much money you lost on gambling?"	won earned lost drank worked	lost won earned drank gained	lost won drank cheated worked	lost won **gamble** earned had	cheated drank won lost earned	lost **gamble** won earned had
BDI	I have no appetite at all anymore. "I have no [MASK] at all anymore."	life hope voice strength memory	**appetite** friends energy family idea	friends energy hope life motivation	money family friends life hope	friends family one life idea	friends energy motivation hope one

Fig. 5. Comparison of the prediction of masked examples from standardized clinical questionnaires. The predicted words are ranked based on the model's estimated probabilities. The correct word is bolded; note that the proposed models often select answers that align with their domain.

Table 7
Mean Reciprocal Rank results over the four reference questionnaires.

	EDE-Q	BDI	NODS-CLiP	SH	Avg
Models					
BERT	0.192	0.244	0.067	0.067	0.142
AnorBERT	**0.574**	**0.452**	0.067	0.150	0.311
DepBERT	0.278	0.432	0.067	0.155	0.233
GambBERT	0.236	0.385	0.136	0.117	0.218
SHBERT	0.236	0.390	0.067	**0.272**	0.241
WholeBERT	0.477	0.406	**0.167**	0.247	**0.324**

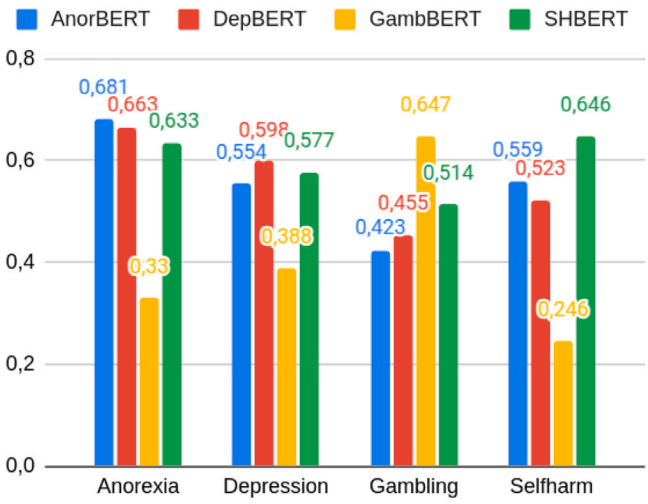

Fig. 6. Average probability assigned by four reference models to the posts published by positive users of Anorexia, Depression, Gambling, and Self-Harm.

words from the domain where the models were specialized. For instance, given the sentence *"I have no [MASK] at all anymore"*, where the word *appetite* was masked, only the anorexia model provided the correct answer. The other models offered logical completions related to friendship, energy, and money. Conversely, in some cases, all models predicted the proper word, albeit with differing priorities. This behavior holds promise for developing tools to assess the prevalence of signs of various mental disorders.

6.1. Exploring the spectrum of mental disorders

There is ample evidence supporting the idea that mental disorders exist on a spectrum and, often, people suffer from multiple disorders [49]. It is not uncommon for an individual to exhibit symptoms of various closely related mental disorders.

To explore the occurrence of symptoms from diverse disorders, we analyzed each eRisk user identified as positive (for each mental disorder) using all the disorder-specialized models. Given the user-generated texts, we obtained the probability of belonging to a specific condition. This analysis can be viewed as a cross-task experiment. For instance, a model trained initially to identify anorexia-related contents is applied to positive users from all four datasets, aiming to uncover symptoms associated with anorexia.

In Table 8, we showcase the posts exhibiting the highest probability as estimated by each model. Observing how the models mark content associated with diverse mental disorders is revealing. For instance, the depression model identifies an anorexia user that reveals comorbidity of depression and anxiety. Similarly, individuals grappling with depression express concerns about eating disorders and articulate a desire to lose weight to fit into smaller clothing sizes. Note also how symptoms indicative of gambling disorders manifest among users coping with different mental health challenges. Some individuals express a keen desire to engage in gambling activities, while others express disdain or aversion toward gambling altogether. This confluence of themes underscores the complex interplay between mental health issues and highlights the nuanced expressions found within online discourse.

To complement this analysis, we calculated the average probabilities computed by these models for the entire set of user posts published by the positive users. This analysis aims to provide a preliminary estimate of the occurrence of symptoms from different disorders in users diagnosed with a specific condition. These prediction averages are reported in Fig. 6. The first exciting thing that we can see is that the gambling model tends to assign low prediction probabilities for anorexia, depression, and self-harm users. This makes sense, as individuals suffering from these three disorders do not necessarily have a gambling problem. It is also interesting to see that anorexia users show traces not only of anorexia but also of depression and self-harm. Likewise, depression users show significant traces of self-harm and anorexia, and self-harm users show substantial traces of depression and anorexia. These results show a strong association among these three disorders.

Table 8
Posts with the highest probabilities estimated by each model. ***Note:** For privacy reasons, texts have been paraphrased using chatGPT.

Anorexia users	
AnorBERT	"I get frustrated sometimes, and I know it's my eating disorder making me think this way. I remind myself that I eat less than I burn and that weight loss has ups and downs. Still, it's discouraging to see the scale higher after a rough day"
DepBERT	"That's awful to hear someone else is struggling too. I know how tough it can be. I'm dealing with atypical anorexia, depression, and anxiety myself"
GambBERT	"Think twice before you gamble. That amount could easily cover two months of college tuition, and I'm struggling to afford it"
SHBERT	"Seven years of therapy have brought progress, but it hasn't been easy. There are moments where it feels like I'm making strides, only to be met with setbacks that leave me feeling like I'm falling back down"
Depression users	
AnorBERT	"I struggle with both restricting my food intake and sometimes overeating. I'm also particular about what I'll eat. Part of this is because I'm uncomfortable with physical development and prefer a smaller body type, especially to fit into clothes designed for thinner people"
DepBERT	"Even though I have a diagnosis of major depressive disorder, I constantly doubt it. It feels like I'm putting on an act of depression and anxiety, even though I know that's not true"
GambBERT	"I'm glad things worked out! Just thinking about gambling and partying made me really anxious. No way, I'd never do that again. I'm so grateful for nature and hiking. They really help me stay on the right track"
SHBERT	"It sounds like social anxiety or maybe a fear of rejection is holding me back. I'm not sure if I even want to be in a relationship. While I enjoy being alone most of the time, being around couples in public triggers deep anxiety in me"
Gambling users	
AnorBERT	"While I value my job, I plan to leave in 6 days. Even though I take care of my health with a good diet, I've been feeling mentally sluggish and confused. I also experience frequent nervousness and fatigue, even after a full night's sleep. It's time for me to address these issues head-on and stop avoiding them"
DepBERT	"I appreciate your help! It's been beneficial. I have a diagnosis, but I also experience severe anxiety. It seems different from social anxiety, though, because I'm not worried about being embarrassed at work or anything like that. What confuses me is figuring out if people are bothered by me"
GambBERT	"This time feels different. I'm done with gambling. It won't be easy, but I'm ready for the fight. I've started over too many times. Gambling is exactly what they say it is - a trap I don't want to be in anymore. Despite a good income for the past 6 years, I'm in debt because of it. Never again"
SHBERT	"Appreciate your advice. It's very helpful. I have a diagnosis of anxiety, but it doesn't quite seem like social anxiety. At work, I'm not worried about being embarrassed or judged by colleagues. There's just this thing where I can't tell if people are getting irritated with me"
Self-harm users	
AnorBERT	"Two weeks into my weight loss journey through diet changes, I decided to ditch soda completely. Now, water is my go-to drink at restaurants and gas stations. What other healthy habits can I incorporate next?"
DepBERT	"It can be tough to open up about mental health struggles. When I first sought help, I was really nervous. What helped me was writing down my symptoms beforehand. Ultimately, having someone there to support you is the most important thing"
GambBERT	"Life's a gamble I can't stomach"
SHBERT	"Crying is really difficult for me. I try to force it out by watching sad movies or listening to sad music, but it doesn't work often. It's important to find healthier ways to cope than hurting yourself. It might seem tempting, but trust me, it won't improve things in the long run".

6.2. Adding interpretability to the detection of signs of mental disorders

While incorporating transformers improves performance compared to other techniques, it also makes model analysis and visualization more challenging. Understanding model behavior is crucial, and in transformer-based models, attention scores in the head modules help identify words or sequences relevant to detection. However, as the number of heads and layers increases, analysis becomes more complex.

To address these challenges, Pardo-Sixtos et al. [50] introduced an interactive visualization tool that graphs attention heads. The tool begins with the [CLS] token and traces back to the most influential tokens from the previous layer (i.e., those with the highest attention values). Such visualization helps to identify the most significant words and sentences in each layer of the transformer, making it easier to analyze key sequences in the text.

The [CLS] token is a special token used by transformers to represent the entire sequence of text. However, it does not indicate which words are most relevant. The graph, instead, illustrates the importance of tokens, ranked from top to bottom based on attention heads. The [CLS] token is the most important, and we can determine which tokens it influences by examining its child nodes in the visualization. For clarity, we typically display only three child nodes. Each child node represents a token that contributes significantly to the meaning of its parent token. Ultimately, the size of the graph is determined by the level of detail you want to analyze. A standard visualization typically includes two or three child nodes per token and has a depth of four or five levels.

For our analysis, we selected the post with the highest probabilities for each task (estimated by the task's most focused model; for example, AnorBERT for Anorexia, see Table 8). We computed attention scores for these posts and used the head module's attention scores to highlight the most relevant parts for classification. Fig. 7 illustrates an example of the generated graphs, helping us better understand the words and contexts crucial to classification.

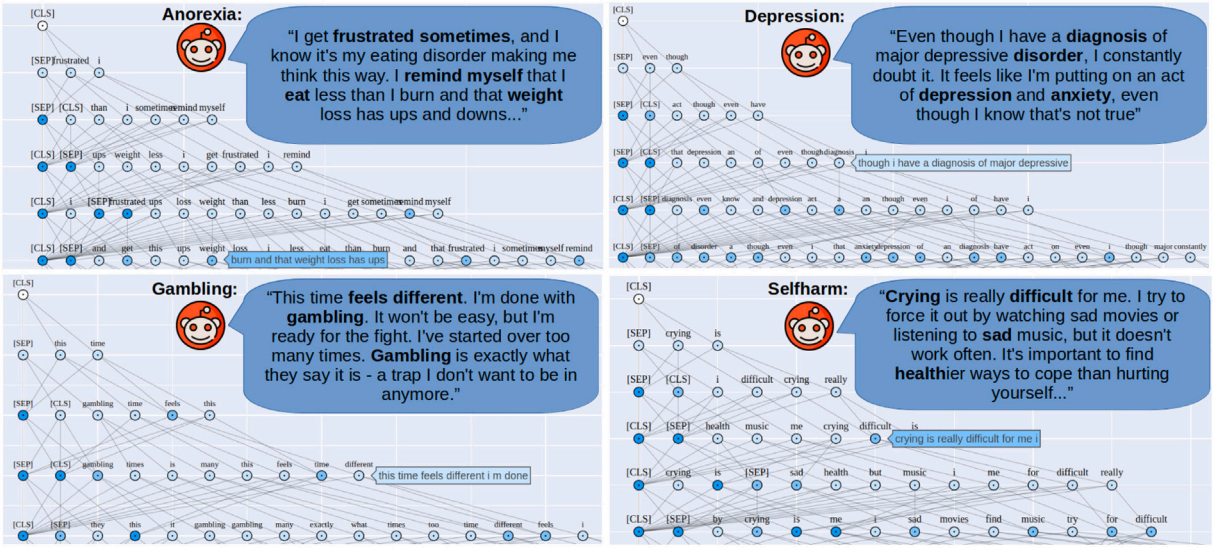

Fig. 7. Graph examples from a user with anorexia (upper-left), depression (upper-right), gambling (lower-left), and self-harm (lower-right). The figure shows the most relevant words in the publication according to the highest attention values.

We can observe that the models highlight key terms associated with mental health struggles, such as "frustrated", "weight loss", "diagnosis", "disorder", "crying", and "difficult". These words play an essential role in the model's understanding of the psychosocial state of the individual (these expressions receive significant attention from multiple tokens). In the anorexia example, the model attends to "burn" and "weight loss has ups", emphasizing the relationship between caloric restriction and weight concerns. For depression, "diagnosis" and "disorder" are key focal points, indicating the model's recognition of clinical language. For gambling, we can observe the focus on explicit words referring to the mental disorder, such as "gambling" and the feeling about it. In self-harm, "crying is really difficult for me" is a focal point, showing how emotional struggle becomes a central theme. The models, therefore, seem to establish meaningful connections between negative emotions and their explanatory context.

7. Cross-platform evaluation

The models reported above were trained on Reddit data. This may introduce a bias due to the platform's unique characteristics (e.g., demographics, usage patterns, or linguistic styles). As a result, their ability to generalize to other social media platforms or mental health contexts – where user behavior may differ – could be limited. To address this concern and broaden our analysis, we explore the classification performance of our models using data from an alternative social media source. Specifically, we have expanded the experimentation to include cross-platform tests, where the models are tested using data from the 2015 ACL Workshop on Computational Linguistics and Clinical Psychology (CLPsych) dataset [51]. This dataset was sourced from X, whose texts (tweets) are largely different from the publications posted on Reddit. For example, the language style and length of the tweets substantially differ from those of Reddit's posts. We set a cross-platform setting where the models trained on Reddit data are tested on the CLPsych dataset. This stringent evaluation scenario allows us to assess the models' robustness and adaptability to other online communities. To put these results in context, we also report results where the models are trained using CLPsych data.

The CLPsych shared task included a classification challenge aimed at categorizing a sample of Twitter users into three mental health groups: (1) users who have self-reported a diagnosis of depression, (2) users who have self-reported a diagnosis of post-traumatic stress disorder (PTSD), and (3) control users who have not reported either condition. For our analysis, we specifically focused on distinguishing between users with self-reported depression and control users. Table 9 provides an overview of the dataset's statistics.

Table 9
CLPsych 2015 dataset used for cross-platform experiments. Statistics of the depression and control groups.

	Depression	Control
# users	477	872
Avg #tweets per user	2373	2286
Avg #words per tweet	13.9	13.8

Table 10
Cross-validation experiments on the CLPsych 2015 dataset. F1, Precision (P), and Recall (R) over the positive class.

	F1	P	R
BERT	.661 ± .032	**.672 ± .103**	.714 ± .070
MentalBERT	.681 ± .021	.627 ± .071	**.753 ± .049**
AnorBERT	.682 ± .014	.643 ± .053	.729 ± .033
DepBERT	**.695 ± .018**	.658 ± .064	.742 ± .041
GambBERT	.680 ± .011	.637 ± .072	.741 ± .063
SHBERT	.685 ± .009	.658 ± .058	.721 ± .047

In our first series of experiments, we trained the six reference models using the CLPsych dataset (see Table 10). Specifically, we conducted 3-fold cross-validation. We can observe that DepBERT achieves the highest F1 score (0.695), indicating that it has a solid balance between precision and recall. MentalBERT yields high recall but lower precision. AnorBERT is inferior to DepBERT in all metrics. Overall, domain-specific adaptations significantly improve results compared to vanilla BERT. Note that BERT yielded the highest precision but was inferior to the other five models regarding recall and F1.

In our second series of experiments, we investigate the transferability of models between Reddit and X: The models trained on Reddit's depression data were evaluated on the CLPsych dataset. Following the methodology of our in- and out-of-domain experiments, each experiment was conducted three times, and we report the average performance and the standard deviation (see Table 11). MentalBERT has the highest F1 score and Recall among the models, and AnorBERT yields the highest precision. The results indicate that performance is lower than that achieved using training data from the same platform. This is a natural consequence of the unavailability of data from the same

Table 11
Cross-platform experiments. F1, Precision (P), and Recall (R) over the positive class in the CLPsych dataset.

	F1	P	R
BERT	.485 ± .034	.455 ± .004	.523 ± .084
MentalBERT	**.506 ± .014**	.455 ± .009	**.570 ± .023**
AnorBERT	.489 ± .024	**.456 ± .003**	.529 ± .062
DepBERT	.487 ± .041	.443 ± .003	.546 ± .104
GambBERT	.486 ± .018	.452 ± .007	.527 ± .043
SHBERT	.490 ± .042	.452 ± .007	.540 ± .105

platform. Still, these cross-platform effectiveness scores are in the range of those reported for same-platform cross-domain experiments (see Table 5). Ideally, we would like to train models using same-platform and same-domain data. However, this luxury is not always available. Our empirical results demonstrate promising potential for transferring models across various platforms and domains. The models achieve F1 scores close to 0.50, suggesting they can identify some signals to detect new, unseen cases from different sources or domains. Refining and training on multiple-source datasets could enhance transferability, closing the gap with same-platform, same-domain models and improving overall performance.

8. Discussion

Our evaluation (as shown in Table 3) suggests that depression represents a more significant challenge for automated detection than the other three disorders. Conditions like anorexia and gambling often exhibit clearer, more easily identifiable symptoms, making them easier to detect using language markers. Self-harm falls somewhere in between, with some outward signs but also with some less obvious cues.

Incorporating adapters into the models' architecture significantly improved their ability to focus on the intricacies of the specific detection tasks. This was particularly beneficial when training with fewer examples, a common situation in mental health research. The adapters enabled the models to extract relevant patterns efficiently from limited evidence. These technological modules are promising techniques to support early risk detection. In fact, adapters have provided key advantages such as computational efficiency, task adaptability (especially with small amounts of data), and little storage space by only needing to save parts of the trained layers for each task.

Training models with examples from other disorders can sometimes yield performance levels comparable to those obtained from models trained with in-domain examples. This demonstrates the potential for cross-domain model transfer. Notably, in cases like anorexia and self-harm, some cross-domain models achieved results close to those produced by the in-domain model. This outcome illustrates the shared psychological concerns among individuals from both groups. On the other hand, the specialized models achieved the best or near-best results even when trained with out-of-domain data. Their strength shows that our task-specific adaptation phase retains relevant information.

We observed an interesting association between the model's training data and performance. Specialized models, which are adapted to specific disorders, when prompted with masked expressions from standardized in-domain clinical instruments tend to provide correct completions. This suggests that our specialization supplies relevant language patterns that enhance the models' accuracy.

It is important to mention that the rapid evolution of Large Language Models (LLMs) presents promising avenues for enhancing digital mental health tools. Recent models, such as MentalLlama [52], have been specifically designed to support users in sensitive mental health contexts, leveraging fine-tuning and safety-aligned dialogue strategies to generate more empathetic and clinically aware responses. Beyond mental health, models like Med-PaLM [53] and GatorTron [54] have demonstrated the potential of LLMs in broader healthcare applications, including medical question answering, clinical decision support, and electronic health record (EHR) analysis. These domain-specific adaptations aim to bridge the gap between general-purpose language understanding and the nuanced requirements of clinical practice. Similarly, Clinical Camel [55] and BioGPT [56] are examples of efforts to align LLM architectures with biomedical knowledge bases, facilitating more accurate and relevant outputs in medical contexts. While these innovations hold great promise, they raise concerns regarding model transparency, reliability, and ethical deployment. In mental health, trust, user safety, and the risk of misinformation require careful oversight. Future research should emphasize fine-tuning LLMs with diverse, high-quality, and clinically validated datasets; integrating human feedback mechanisms; and developing rigorous evaluation metrics aligned with psychological and psychiatric outcomes. Close interdisciplinary collaboration between AI researchers, clinicians, ethicists, and patients will be essential to advance the role of LLMs in mental healthcare.

Note also that detecting patterns associated with mental disorders can play a crucial role in supporting mental health professionals. Our technology could be exploited to identify emerging trends and generate real-time alerts that help clinicians or relevant institutions. For example, this could help to recognize shifts in mental health risks within specific populations. These new forms of computer-based screening could assist in designing early intervention strategies, informing public health campaigns, and enhancing clinical decision-making through risk assessment dashboards. Additionally, our framework could be integrated into healthcare systems to provide valuable insights into social media behavior. For instance, flagged risk signals – such as language patterns indicative of distress or suicidal ideation – could serve as supplementary data for psychiatric evaluations or risk assessments. By leveraging this information, mental health professionals can better understand trends and tailor their campaigns or interventions accordingly. Another possible line of exploitation would be to transfer the classifiers built from massive web data for tracking and monitoring patient-generated texts (e.g., processing anonymized written essays in Psychological Therapy to detect the presence of different symptoms).

9. Limitations and ethical statement

Below, we outline the limitations of our study:

1. Data Collection Challenges and Privacy Concerns: The datasets publicly accessible for detecting signs of mental disorders are often constrained to limited sample sizes. This is due to the complexities of data collection and privacy considerations.

2. Observational Nature of Research: Our study adopts an observational approach. We lack access to personal and psychological data from these subjects, which is typically integral to risk assessment investigations in the clinical domain.

3. Inherent Bias from Data Source: A bias unavoidably arises from the data source, as only users engaged with social media platforms, particularly Reddit, were included in the study. Consequently, relevant segments of the population, such as the elderly or individuals who consciously refrain from maintaining online profiles (or opt for privacy settings), remain beyond the scope of our monitoring efforts.

Regarding ethical considerations, we acknowledge the complexities of analyzing social media content. Exploring online data often raises concerns regarding privacy and ethics. **Privacy and confidentiality:** Mental health data are highly sensitive and personal; some risks include data breaches, unauthorized access, and potentially re-identifying individuals in anonymized datasets. This could lead to privacy violations, discrimination, or exploitation for marketing purposes, causing harm to the user's life. Researchers should adhere to their Institutional Review Boards (IRB) and avoid the misuse of models and data. **Data protection:** Social networking data may not be adequately protected, which can lead to breaches of privacy and confidentiality. **Dependency:** Other potential risks are associated with overreliance on automated systems.

Models may produce false positives or negatives, leading to inaccurate predictions. Overreliance on these systems without human oversight could harm people. As researchers, we should also avoid reinforcing stereotypes. **Bias and discrimination:** Sometimes, the training data could be biased, and the model may disproportionately flag certain groups (e.g., racial minorities and LGBTQ+ individuals) as having mental health issues, perpetuating harmful stereotypes. We should regularly test the biases of the models and ensure they perform equitably across different demographic groups when possible. **Lack of context**: Social network data can lack context, leading to misinterpretations or incomplete interpretations of information.

Accordingly, for this study, we only used publicly available data and existing research collections whose terms and conditions explicitly protect against misuse. We did not contact any social media users. Furthermore, this research project was formally approved by our IRB review (USC 65/2024). Moreover, the datasets solely contain public user interactions, and we meticulously adhered to these text collections' terms of use and user agreements. Additionally, it is essential to note that the datasets have been anonymized to safeguard privacy. While individuals may share posts publicly, they may not anticipate their content reaching a broad audience. Consequently, we have paraphrased the excerpts showcased in this paper to respect privacy and confidentiality. We must emphasize our commitment to using this research to improve society. Specifically, we aim to extract insights to improve mental health diagnoses and assist individuals with mental disorders. Our effort is based on a desire to influence society's well-being positively.

10. Conclusions and future work

In this study, we addressed a pressing global issue: the prevalence of mental disorders. In this context, we aim to drive positive advancements in automated methods for screening these disorders. To achieve this, we have designed and implemented a framework focused on identifying early indicators of anorexia, depression, gambling, and self-harm in social media content.

Our efforts have encompassed the entire data processing pipeline, from data acquisition in relevant tasks to domain adaptation and task-specific adjustments. We also leveraged advanced technological techniques, such as adapters, to improve the efficiency and effectiveness of the resulting neural architectures.

Our findings indicate that our adaptation approach is solid for detecting relevant language markers of mental disorders, surpassing baseline models. Furthermore, it demonstrates competitive performance comparable to leading early-risk algorithms. Additionally, we offered a comprehensive analysis of our models, unveiling the contextual nuances captured and providing a deeper insight into individuals' expressions of concern.

In future work, we aim to utilize additional lexical resources to improve the domain adaptation process and leverage clinical data to refine specialized language models. We are also interested in deploying this tool to extract relevant linguistic cues associated with mental disorders, thereby supporting psychologists in screening for manifestations of mental health traits. Moreover, we aspire to expand our investigation to other languages, as we plan to develop corpora and models for these languages in the near future.

CRediT authorship contribution statement

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

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 article.

Acknowledgments

We thank CONAHCYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies. 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).

Appendix. Subreddits used for data collection

In the following list, we detail the different subreddits selected for each psychological disorder:

- **Anorexia:** "AskPsychiatry", "mentalhealth", "mentalillness", "bingeeating", "bulimia", "Health", "eating_disorders", "EatingDisorderHope", "EatingDisorders", "emetophobia", "fuckeatingdisorders", "EDAnonymous", "MaleEatingDisorders", "EatingDisorders_male", "BingeEatingDisorder", "eatingdisordered", "AnorexiaNervosa", "eatingdisorderstories".
- **Depression:** "depression", "depressionregimens", "sad", "depressed", "depression_help", "depressionregimens", "DepressionJournals", "dysthymia", "AnxietyDepression", "antidepressants", "PsychiatricFreedom", "lexapro", "prozac", "zoloft", "SSRIs", "AskPsychiatry", "mentalhealth", "mentalillness", "MMFB", "depression_partners", "AdultDepression", "DepressionNests", "depressionregimens", "SuicideWatch", "Tackle_depression", "Postpartum_Depression", "DepressionBuddies", "Depression Cleaning", "DepressionMusic", "DepressionResearch" , "DepressionPoems", "GFD", "EOOD".
- **Gambling:** "gambling", "GamblingAddiction", "GamblingStrategies", "problemgambling", "wallstreetbets", "GamblingSites", "OnlineCasinoGambling", "CryptoGambling", "SimpleOnline Gambling", "RealGambling", "GamblingBonuses", "onlinegambling", "sportsgambling", "BitcoinGambling", "online_gambling", "DailyGambling", "poker", "GamblingHall", "SMARTRecovery", "GolfGambling", "RedditGambling", "doge_gambling", "SolutionGambling", "OnlineCryptoGambling", "mentalhealth", "mentalillness".
- **Self-harm:** "AskPsychiatry", "mentalhealth", "mentalillness", "selfharmhelpers", "Selfharmscarsrecovery", "selfharmersunite", "SuicideWatch", "selfharm", "black_selfharm", "selfharmteens", "AdultSelfHarm", "StopSelfHarm", "PsychiatricFreedom", "selfhelp", "sad", "selfharmcope", "Cutters".

References

[1] World Health Organization W. Mental health: Fact sheet. 2019, https://www.who.int/en/health-topics/noncommunicable-diseases/mental-health.
[2] Kessler R, Bromet E, Jonge P, Shahly V, Marsha. The burden of depressive illness. Public Heal Perspect Depressive Disord 2017;40–66.
[3] Foy C. Are mental disorders increasing over time?. 2021, [Accessed 23 February 2024] http://tinyurl.com/3pts3rec.
[4] Ríssola EA, Losada DE, Crestani F. A survey of computational methods for online mental state assessment on social media. ACM Trans Comput Heal 2021;2(2). http://dx.doi.org/10.1145/3437259.
[5] Crestani F, Losada DE, Parapar J. Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project. Studies in Computational Intelligence XII, first ed. Englewood Cliffs, NJ: Springer Verlag; 2022, p. 328. http://dx.doi.org/10.1007/978-3-031-04431-1.
[6] Renteria-Rodriguez ME. Salud mental en Mexico. In: NOTA- INCyTU NúMERO 007. 2018.

[7] Howard J, Ruder S. 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). Melbourne, Australia: Association for Computational Linguistics; 2018, p. 328–39. http://dx.doi.org/10.18653/v1/P18-1031, URL https://aclanthology.org/P18-1031.

[8] Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, Smith NA. Don't stop pretraining: Adapt language models to domains and tasks. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th annual meeting of the association for computational linguistics. Online: Association for Computational Linguistics; 2020, p. 8342–60. http://dx.doi.org/10.18653/v1/2020.acl-main.740, URL https://aclanthology.org/2020.acl-main.740.

[9] Pfeiffer J, Rücklé A, Poth C, Kamath A, Vulić I, Ruder S, Cho K, Gurevych I. AdapterHub: A framework for adapting transformers. In: Liu Q, Schlangen D, editors. Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Online: Association for Computational Linguistics; 2020, p. 46–54. http://dx.doi.org/10.18653/v1/2020.emnlp-demos.7, URL https://aclanthology.org/2020.emnlp-demos.7.

[10] De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: Proceedings of the 7th international AAAI conference on weblogs and social media. 2013, p. 128–37.

[11] De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations.. In: Proceedings of the 5th annual ACM web science conference. 2013, p. 47–56.

[12] Wang T, Brede M, Ianni A, Mentzakis E. Detecting and characterizing eating-disorder communities on social media. In: Proceedings of the Tenth ACM International conference on web search and data mining. 2017, p. 91–100.

[13] Ortega-Mendoza RM, Hernández-Farías DI, y Gómez MM, Villaseñor-Pineda L. Revealing traces of depression through personal statements analysis in social media. Artif Intell Med 2022;123:102202. http://dx.doi.org/10.1016/j.artmed.2021.102202.

[14] Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H. Recognizing depression from twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Fac- tors in Computing Systems. 2015, p. 3187–96.

[15] Schwartz H, Eichstaedt J, Kern M, Park G, Sap M, Stillwell D, Kosinski M, Ungar L. Towards assessing changes in degree of depression through facebook. Proc Work Comput Linguist Clin Psychol: From Linguist Signal To Clin Real 2014;118–25.

[16] Trotzek M, Koitka S, Friedrich C. Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia. In: Proceedings of the 9th international con- ference of the CLEF association. Avignon, France: CLEF 2018; 2018.

[17] Coppersmith G, Harman C, Dredze M. Measuring post traumatic stress disorder in Twitter. In: Proceedings of the eighth in- ternational AAAI conference on weblogs and social media. 2014, p. 579–82.

[18] Trifan A, Oliveira J. BioInfo@uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders. In: Proceedings of the 10th International Conference of the CLEF Association,. Lugano, Switzerland: CLEF 2019; 2019.

[19] Van Rijen P, Teodoro D, Naderi N, Mottin L, Knafou J, Jeffryes M, Ruch P. A data-driven approach for measuring the severity of the signs of depression using reddit posts. In: Proceedings of the 10th International Conference of the CLEF Associa- tion. Lugano, Switzerland: CLEF 2019; 2019.

[20] Helmy A, Nassar R, Ramdan N. Depression detection for twitter users using sentiment analysis in english and arabic tweets. Artif Intell Med 2024;147:102716. http://dx.doi.org/10.1016/j.artmed.2023.102716.

[21] Bertl M, Bignoumba N, Ross P, Yahia SB, Draheim D. Evaluation of deep learning-based depression detection using medical claims data. Artif Intell Med 2024;147:102745. http://dx.doi.org/10.1016/j.artmed.2023.102745, URL https://www.sciencedirect.com/science/article/pii/S0933365723002592.

[22] Pérez A, Parapar J, Barreiro Á. Automatic depression score estimation with word embedding models. Artif Intell Med 2022;132:102380. http://dx.doi.org/10.1016/j.artmed.2022.102380.

[23] Wu C-S, Chen C-H, Su C-H, Chien Y-L, Dai H-J, Chen H-H. Augmenting DSM-5 diagnostic criteria with self-attention-based bilstm models for psychiatric diagnosis. Artif Intell Med 2023;136:102488. http://dx.doi.org/10.1016/j.artmed.2023.102488, URL https://www.sciencedirect.com/science/article/pii/S0933365723000027.

[24] Masood R. Adapting models for the case of early risk prediction on the internet. Adv Inf Retr ECIR 2019. Lect Notes Comput Sci Vol 11438. 2019;353–8.

[25] Hossain MM, Hossain MS, Mridha MF, Safran M, Alfarhood S. Multi task opinion enhanced hybrid BERT model for mental health analysis. Sci Rep 2025;15(1):3332. http://dx.doi.org/10.1038/s41598-025-86124-6.

[26] Pourkeyvan A, Safa R, Sorourkhah A. Harnessing the power of hugging face transformers for predicting mental health disorders in social networks. IEEE Access 2024;12:28025–35. http://dx.doi.org/10.1109/ACCESS.2024.3366653.

[27] Chopra S, Agarwal P, Ahmed J, Biswas SS, Obaid AJ. Roberta and BERT: Revolutionizing mental healthcare through natural language. SN Comput Sci 2024;5(7):889. http://dx.doi.org/10.1007/s42979-024-03202-8.

[28] Gaurav A, Gupta BB, Chui KT. BERT based model for robust mental health analysis in clinical informatics. In: 2024 21st international joint conference on computer science and software engineering. JCSSE, 2024, p. 153–60. http://dx.doi.org/10.1109/JCSSE61278.2024.10613729.

[29] Martinez-Romo J, Araujo L, Reneses B. Guardian-BERT: Early detection of self-injury and suicidal signs with language technologies in electronic health reports. Comput Biol Med 2025;186:109701. http://dx.doi.org/10.1016/j.compbiomed.2025.109701, URL https://www.sciencedirect.com/science/article/pii/S0010482525000514.

[30] Zihan L, Yan X, Tiezheng Y, Wenliang D, Ziwei J, Samuel C, Andrea M, Pascale F. CrossNER: Evaluating cross-domain named entity recognition. AAAI Conf Artif Intell 2021;35:13452–60.

[31] Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019, p. 3615–20. http://dx.doi.org/10.18653/v1/D19-1371, URL https://aclanthology.org/D19-1371.

[32] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019;36(4):1234–40. http://dx.doi.org/10.1093/bioinformatics/btz682.

[33] Aragon M, Lopez Monroy AP, Gonzalez L, Losada DE, Montes M. DisorBERT: A double domain adaptation model for detecting signs of mental disorders in social media. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). Toronto, Canada: Association for Computational Linguistics; 2023, p. 15305–18. http://dx.doi.org/10.18653/v1/2023.acl-long.853, URL https://aclanthology.org/2023.acl-long.853.

[34] Houlsby N, Giurgiu A, Jastrzebski S, Morrone B, de Laroussilhe Q, Gesmundo A, Attariyan M, Gelly S. Parameter-efficient transfer learning for NLP. 2019, ArXiv, arXiv:1902.00751 URL https://api.semanticscholar.org/CorpusID:59599816.

[35] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Brew J. Fine-tuning a masked language model. 2022, URL https://huggingface.co/course/chapter7/3?fw=pt.

[36] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. 2015, arXiv preprint arXiv:1508.07909.

[37] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019, p. 4171–86. http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

[38] He J, Zhou C, Ma X, Berg-Kirkpatrick T, Neubig G. Towards a unified view of parameter-efficient transfer learning. 2021, ArXiv, arXiv:2110.04366 URL https://api.semanticscholar.org/CorpusID:238583580.

[39] Losada DE, Crestani F, Parapar J. Overview of erisk 2019 early risk prediction on the internet. In: Experimental IR meets multilinguality, multimodality, and interaction. Cham: Springer International Publishing; 2019, p. 340–57. http://dx.doi.org/10.1007/978-3-030-28577-7_27.

[40] Losada DE, Crestani F, Parapar J. Overview of erisk 2020: Early risk prediction on the internet. In: Experimental IR meets multilinguality, multimodality, and interaction. Cham: Springer International Publishing; 2020, p. 272–87. http://dx.doi.org/10.1007/978-3-030-58219-7_20.

[41] Parapar J, Martín-Rodilla P, Losada DE, Crestani F. Overview of erisk 2021: Early risk prediction on the internet. In: Experimental IR meets multilinguality, multimodality, and interaction. Cham: Springer International Publishing; 2021, p. 324–44.

[42] Parapar J, Martín-Rodilla P, Losada DE, Crestani F. Overview of eRisk 2022: Early risk prediction on the internet. In: Experimental IR meets multilinguality, multimodality, and interaction. Cham: Springer International Publishing; 2022, p. 233–56.

[43] Parapar J, Martín-Rodilla P, Losada DE, Crestani F. Overview of eRisk 2023: Early risk prediction on the internet. In: Experimental IR meets multilinguality, multimodality, and interaction. Cham: Springer Nature Switzerland; 2023, p. 294–315.

[44] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems 32. Curran Associates, Inc.; 2019, p. 8024–35.

[45] Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: Publicly available pretrained language models for mental healthcare. In: Proceedings of the thirteenth language resources and evaluation conference. Marseille, France: European Language Resources Association; 2022, p. 7184–90, URL https://aclanthology.org/2022.lrec-1.778.

[46] Beck AT, Steer RA, Carbin MG. Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. Clin Psychol Rev 1988;8(1):77–100.

[47] Fairburn CG, Beglin SJ. Assessment of eating disorders: Interview or self-report questionnaire? Int J Eat Disord 1994;16(4):363–70. http://dx.doi.org/10.1002/1098-108X(199412)16:4<363::AID-EAT2260160405>3.0.CO;2-#.

[48] Xian H, Shah K, Phillips S, Scherrer J, Volberg R, Eisen S. Association of cognitive distortions with problem and pathological gambling in adult male twins. Psychiatry Res 2008;160(3). http://dx.doi.org/10.1016/j.psychres.2007.08.007.

[49] Adam D. Mental health: On the spectrum. Nature 2013;416–8.

[50] Pardo-Sixtos LF, López-Monroy AP, Shafaei M, Solorio T. Hierarchical attention and transformers for automatic movie rating. Expert Syst Appl 2022;209:118164. http://dx.doi.org/10.1016/j.eswa.2022.118164, URL https://www.sciencedirect.com/science/article/pii/S0957417422013240.

[51] Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. 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, p. 31–9.

[52] Yang K, Zhang T, Kuang Z, Xie Q, Ananiadou S. MentalLLaMA: Interpretable mental health analysis on social media with large language models. 2023, arXiv preprint arXiv:2309.13567.

[53] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera y Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80. http://dx.doi.org/10.1038/s41586-023-06291-2.

[54] Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, Compas C, Martin C, Flores MG, Zhang Y, Magoc T, Harle CA, Lipori G, Mitchell DA, Hogan WR, Shenkman EA, Bian J, Wu Y. GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records. 2022, arXiv:2203.03540 URL https://arxiv.org/abs/2203.03540.

[55] Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. 2023, arXiv:2305.12031 URL https://arxiv.org/abs/2305.12031.

[56] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y. Biogpt: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23(6):bbac409. http://dx.doi.org/10.1093/bib/bbac409, arXiv:https://academic.oup.com/bib/article-pdf/23/6/bbac409/47144271/bbac409.pdf.