Delving into the Depths: Evaluating Depression Severity through BDI-biased Summaries

Mario Ezra Aragón$^{\alpha}$**, Javier Parapar**$^{\beta}$**, David E. Losada**$^{\alpha}$
$^{\alpha}$ Centro Singular de Investigación en Tecnoloxias Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
$^{\beta}$ Universidade da Coruña, Spain
{ezra.aragon,david.losada}@usc.es, javier.parapar@udc.es

Abstract

Depression is a global concern suffered by millions of people, significantly impacting their thoughts and behavior. Over the years, heightened awareness, spurred by health campaigns and other initiatives, has driven the study of this disorder using data collected from social media platforms. In our research, we aim to gauge the severity of symptoms related to depression among social media users. The ultimate goal is to estimate the user's responses to a well-known standardized psychological questionnaire, the Beck Depression Inventory-II (BDI). This is a 21-question multiple-choice self-report inventory that covers multiple topics about how the subject has been feeling. Mining users' social media interactions and understanding psychological states represents a challenging goal. To that end, we present here an approach based on search and summarization that extracts multiple BDI-biased summaries from the thread of users' publications. We also leverage a robust large language model to estimate the potential answer for each BDI item. Our method involves several steps. First, we employ a search strategy based on sentence similarity to obtain pertinent extracts related to each topic in the BDI questionnaire. Next, we compile summaries of the content of these groups of extracts. Last, we exploit chatGPT to respond to the 21 BDI questions, using the summaries as contextual information in the prompt. Our model has undergone rigorous evaluation across various depression datasets, yielding encouraging results. The experimental report includes a comparison against an assessment done by expert humans and competes favorably with state-of-the-art methods.

1 Introduction

Nowadays, numerous individuals in the world suffer from diverse mental conditions that disrupt their cognition and conduct and, ultimately, represent a detriment to their quality of life (Kessler et al., 2017). As an illustration, depression stands out as one of the most prevalent mental disorders, positioning itself as a primary catalyst for suicidal tendencies (Mathers and Loncar, 2006). A major depressive disorder is a significant medical condition that has adverse effects on emotions, thoughts, and behaviors. Depression induces feelings of sadness and a diminished interest in previously enjoyable activities. This condition can result in various emotional and physical challenges, impacting one's ability to perform effectively both in the workplace and at home (APA, 2020). Currently, only approximately 20% of those afflicted receive necessary early intervention, with a significant proportion of mental health expenditures allocated to the maintenance of psychiatric institutions as opposed to activities encompassing detection, prevention, and recovery (Renteria-Rodriguez, 2018). Given these circumstances, there exists an urgent need to design effective approaches for the early detection of depression, aiming to avoid harm to individuals suffering from this condition.

The ubiquity of social media data has paved the way for data-driven research in the field of mental health analysis (Ríssola et al., 2021; Skaik and Inkpen, 2020). A significant portion of individuals conduct the bulk of their social interactions within the digital realm crafted by social media platforms such as Facebook, Twitter, Reddit, and Instagram. Nowadays, researchers have access to extensive corpora of online dialogues on diverse topics. This wealth of data holds particular significance in medicine, where progress in our understanding of mental health could directly contribute to life-saving quality-of-life measures and improvements.

Exploiting public interactions offers a valuable avenue for comprehending depression, thereby amplifying the potential to identify individuals displaying depressive indicators and facilitating professional intervention (Ríssola et al., 2021; Crestani

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 12–22
March 21, 2024

et al., 2022a). Diverse techniques rooted in Natural Language Processing (NLP), Text Classification (TC), and Information Retrieval (IR) have been employed to discern signs of depression, with a particular focus on linguistic and sentiment analysis (Crestani et al., 2022b). However, most of the existing studies have been confined to distinguishing between a depression group and a control group (two-class classification) and provide no further explanation or explicit standardized signs that health professionals can analyze. Furthermore, conventional strategies have demonstrated their effectiveness in detecting depressive individuals based on their textual interactions (Velupillai et al., 2019), but they heavily rely on the intricate process of feature engineering (e.g., by extracting optimal user attributes that reflect the subject's feelings and psychological state). However, the NLP landscape has radically evolved in recent years, with the ascent of Large Language Models (LLMs). New models, such as chatGPT, have gained immense popularity due to their capacity to deliver zero-shot and few-shot predictions across diverse tasks[1]. This ability stems from the LLMs' augmented scale, with a huge number of parameters that inherently empower them to encapsulate the subtleties inherent in massive amounts of textual data. This becomes particularly pivotal when confronting linguistic data, given the inherent variance in word significance dependent on the context. To properly exploit current LLMs to support BDI-based screening, the parametric knowledge of the LLM, which provides a sophisticated understanding of human language, needs to be enriched with user-specific interactions related to standardized depression symptoms. This is precisely the main goal of our research. More specifically, this study designs effective search strategies to mine BDI-biased summaries from the users' posting history and proposes the utilization of LLMs for quantifying levels of depression.

Our approach can be regarded as a retrieval-then-read method (Zhu et al., 2021) that augments the LLM knowledge with personalized BDI-biased summaries built for each category of the BDI-II questionnaire. BDI (Beck et al., 1961) is a recognized psychological instrument designed to assess the manifestation of 21 depressive symptoms, such as sadness, pessimism, or loss of energy. We can summarize our contributions as follows:

1. We extract relevant sentences related to different topics of depression to measure the severity of signs of depression among social media users.

2. We explore the use of summaries for each group of sentences to provide an estimated answer to each question in the BDI questionnaire.

3. We empirically evaluate the proposed model and provide quantitative and qualitative evidence of its robustness for the evaluation of depression levels. This includes a comparison against a human expert (trained psychologist), who was also presented with the BDI-biased summaries.

2 Related Work

The examination of public mental health via social media has experienced significant growth in recent years (Ríssola et al., 2021; Skaik and Inkpen, 2020; Guntuku et al., 2017). Recent research has focused on depressive symptom detection to enhance mental health models, highlighting their potential to enhance performance, general applicability, and interpretability (Crestani et al., 2022a; Parapar et al., 2023). For instance, in Nguyen et al. (2022), the authors introduced methods for identifying depression that incorporate various levels of constraints based on the symptoms outlined in the PHQ9 questionnaire, a tool used by clinicians for screening depression. Their experiments, conducted across three social media datasets, revealed that their model can adapt to unfamiliar data, surpassing a conventional BERT-based approach. Another study (Pérez et al., 2022a) presented an approach for automatically gauging the severity of depression in social media users. This research team tackled the task of quantifying the intensity of depression indicators and explored using neural language models to capture different facets of a user's writings. They presented two alternative methodologies to assess the sensitivity of symptoms in terms of the user's willingness to openly discuss them. The first method relies on global language patterns from the user's posts, while the second method seeks direct mentions of symptom-related concerns. Both techniques led to automatic estimates of the overall BDI-II score. Furthermore, in Pérez et al. (2022b), an efficient semantic pipeline was introduced for evaluating depression severity

[1] OpenAI. (2023). chatGPT. https://chat.openai.com/chat

in individuals based on their social media content. The authors selected a sample of user sentences to create semantic rankings. The approach was supported by a reference index of training sentences that correspond to depressive symptoms and severity levels. Subsequently, they employed the sentences derived from these rankings as evidence for predicting the severity of symptoms in users.

In a different direction, Zhang et al. (2022) introduced a method for screening risky posts guided by psychiatric scales. This method identified posts that exhibit risk factors associated with the dimensions outlined in clinical depression scales, providing a basis for a comprehensible diagnosis. To enhance the transparency of predictions, this team proposed a Hierarchical Attentional Network integrated with BERT, known as HAN-BERT.

In recent years, with the proliferation of Large Language Models (LLMs), there has been a response to the limitations observed in psychological knowledge by developing specialized language models that offer improved accuracy in providing psychological advice (Li et al., 2023). Such endeavors have sparked our interest in exploring the potential of LLMs to respond to questionnaires related to depression symptoms and compare them with the assessment done by an expert psychologist. Our approach can be seen as a novel application of retrieve-then-read methods for LLMs (Nishida et al., 2018; Izacard and Grave, 2021), where the parametric LLM model is conditioned by personalized summaries for each user.

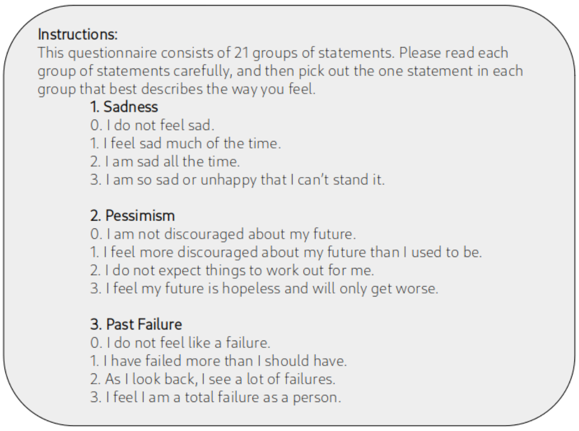
3 Proposed Approach

The objective of this research consists of estimating the level of depression from a thread of users' posts (Losada et al., 2019). Additionally, we contrast our estimates with the answers provided by an expert psychologist, who is also presented with user-level evidence mined from social media. To that end, we summarize the post history of each user, and our model estimates the response to each BDI item based on the evidence found in BDI-specific summaries. The approach consists of three main steps:

1. Extraction of relevant sentences for each of the 21 topics of the BDI questionnaire.

2. Generation of a BDI-biased summary from each group of sentences.

3. Estimation of the response to each BDI question using a large language model.



Instructions:
This questionnaire consists of 21 groups of statements. Please read each group of statements carefully, and then pick out the one statement in each group that best describes the way you feel.

1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

Figure 1: Beck's Depression Inventory. This questionnaire consists of 21 items related to various symptoms of depression. The figure shows three examples.

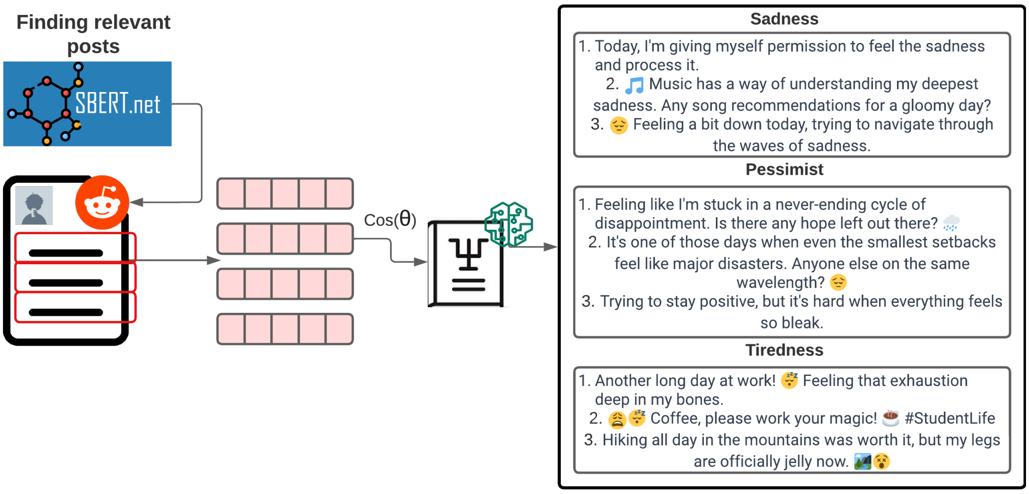
The BDI (Beck et al., 1961) consists of a series of multiple-choice questions or statements about various symptoms and attitudes related to depression (see Figure 1). Respondents are asked to select the statement that best describes their feelings. Each item in the BDI is assigned a score, ranging from 0 to 3, with higher scores indicating more severe symptoms[2]. An overall depression score is obtained by summing the scores for all items. The higher the total score, the more severe the depression is considered to be. This psychometric assessment has been widely employed as a dependable method for gathering high-quality data from various sources, including online sources (Choudhury et al., 2013; Guntuku et al., 2017).

3.1 Extraction of relevant sentences for each BDI item

The first step involves the extraction of relevant sentences for each topic in the BDI questionnaire. First, we convert each question of the BDI to an embedding representation using sentenceBERT (Reimers and Gurevych, 2019), a modification of the pre-trained BERT that yields semantically meaningful sentence embeddings. For each topic, we take the possible responses and the title of the BDI item to create embedding representations. The objective is to create a dictionary of embeddings that represents the BDI questionnaire.

For each social media user, we segment his thread of publications and measure the similarity between the user's sentences and the embeddings

[2] https://www.ismanet.org/doctoryourspirit/pdfs/Beck-Depression-Inventory-BDI.pdf



Sadness

1. Today, I'm giving myself permission to feel the sadness and process it.
2. 🎵 Music has a way of understanding my deepest sadness. Any song recommendations for a gloomy day?
3. 😔 Feeling a bit down today, trying to navigate through the waves of sadness.

Pessimist

1. Feeling like I'm stuck in a never-ending cycle of disappointment. Is there any hope left out there? 🌧️
2. It's one of those days when even the smallest setbacks feel like major disasters. Anyone else on the same wavelength? 😔
3. Trying to stay positive, but it's hard when everything feels so bleak.

Tiredness

1. Another long day at work! 😴 Feeling that exhaustion deep in my bones.
2. 😩😴 Coffee, please work your magic! ☕ #StudentLife
3. Hiking all day in the mountains was worth it, but my legs are officially jelly now. 🏞️😵

Figure 2: Searching for relevant sentences in the user's history. The sentences are grouped by BDI topic and the resulting set of sentences might contain sentences with different polarity.

in the dictionary by applying cosine similarity between each pair of vectors.

Finally, to select candidate sentences for each BDI topic, we empirically establish a threshold of 0.4 and choose the sentences whose similarity is higher than this threshold. A user's sentence is selected for the BDI item as long as it is similar to at least one of the embeddings in the BDI item's group. Figure 2 illustrates this selection process. We can see how the method seeks sentences that are on-topic concerning each BDI item. Note that it can select on-topic sentences with a negative or positive valence. For example, the sentence "Hiking all day in the mountains was worth it, but my legs are officially jelly now" is relevant to tiredness but, in this case, describes a pleasing activity done outdoors.

3.2 Generating summaries of the extracted sentences

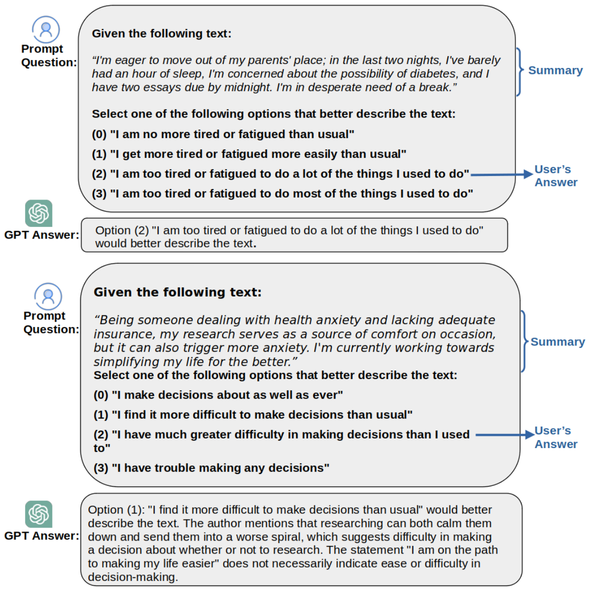
The next step is to create a summary for each group of selected sentences. The idea is to present the LLM with condensed information for each BDI item. LLMs typically have a token input limit and, thus, we cannot feed them with an arbitrarily large sequence of sentences. Restricting the analysis to succinct summaries is also beneficial for reducing the effort required from the human psychologist in her assessment.

For summarization, we used BART, a denoising autoencoder for sequence-to-sequence models (Lewis et al., 2020). It uses a standard Transformer-based architecture, which can be seen as a generalization of both BERT (due to the bidirectional encoder) and GPT (with the left-to-right decoder). More specifically, we employed the model that was obtained by fine-tuning BART on the SAMSum dataset[3]. The SAMSum dataset contains about 16k messenger-like conversations and summaries. The conversations were created and written down by linguists fluent in English. The style and register are diversified, and conversations could be informal, semi-formal, or formal, and they may contain slang words, emoticons, and typos. This represents a language style that is similar to the one in Reddit publications. With the trained model, for each topic of the BDI, we fed the group of relevant posts to BART and generated a summary.

3.3 Estimating the responses of the BDI questionnaire

The last step consists of answering the BDI questionnaire for each user using the generated summaries. To that end, we prompted chatGPT (for these experiments we used the GPT, versions 3.5-turbo-0613 and 4) with the summary and proper instructions. For each user, the prompted questions were processed within a continuous chat. The answer to each BDI question was obtained by parsing the LLM's output. In Figure 3, we can see two examples of these prompts. chatGPT is instructed to select the option that best describes the user's text (the corresponding summary). The options are the answers to each topic within the BDI, ranging from 0 to 3. For illustrative purposes, we added to the figure the answer the user selected for that question (marked with a blue arrow). At the bottom, we can

[3] https://huggingface.co/datasets/samsum



Given the following text:

"I'm eager to move out of my parents' place; in the last two nights, I've barely had an hour of sleep, I'm concerned about the possibility of diabetes, and I have two essays due by midnight. I'm in desperate need of a break."

Select one of the following options that better describe the text:

(0) "I am no more tired or fatigued than usual"

(1) "I get more tired or fatigued more easily than usual"

(2) "I am too tired or fatigued to do a lot of the things I used to do"

(3) "I am too tired or fatigued to do most of the things I used to do"

Option (2) "I am too tired or fatigued to do a lot of the things I used to do" would better describe the text.

Given the following text:

"Being someone dealing with health anxiety and lacking adequate insurance, my research serves as a source of comfort on occasion, but it can also trigger more anxiety. I'm currently working towards simplifying my life for the better."

Select one of the following options that better describe the text:

(0) "I make decisions about as well as ever"

(1) "I find it more difficult to make decisions than usual"

(2) "I have much greater difficulty in making decisions than I used to"

(3) "I have trouble making any decisions"

Option (1): "I find it more difficult to make decisions than usual" would better describe the text. The author mentions that researching can both calm them down and send them into a worse spiral, which suggests difficulty in making a decision about whether or not to research. The statement "I am on the path to making my life easier" does not necessarily indicate ease or difficulty in decision-making.

Figure 3: Prompt questions examples. The text represents the summary of the posts and the blue arrow is the answer selected by the user.

also see the answer predicted by chatGPT and a description of why the model chose that answer. We can see that the model can generally approximate the answers by having the right context in the summary. If the model provided an answer that was not in the range of the possible responses then the output was taken as 0, which represents the absence of negative signs for the corresponding BDI item.

Additionally, to contrast the automatic estimates and performance of the large language model, we also gave the summaries to an expert in the field and asked her to provide her estimated responses to the questionnaires.

4 Experimental Settings

4.1 Data collections

For evaluation, we employed the data sets from the eRisk 2019-2021 evaluation tasks (Losada et al., 2019, 2020; Parapar et al., 2021) on measuring the severity of the signs of depression. The task consists of estimating the level of the 21 standardized depression symptoms based on a thread of user posts. The collection contains a self-report BDI inventory filled by each user in the collection and the users' publications on Reddit. The 2019 dataset consists of 20 users, while 2020 and 2021 have 70 and 80 users respectively. This dataset contains an average number of posts per user of 518 and an average number of words for each post of 40. To select the sample, the creators of these datasets asked online users (particularly within certain mental health subreddits) to fill out the BDI questionnaire and to give consent to analyze their public interactions. These BDI questionnaires act as the ground truth to contrast the questionnaires filled by the system or by the health expert.

Pre-processing: We performed a simple pre-processing on the user-generated texts by lowercasing all words and removing special characters like URLs, emoticons, and hashtags.

4.2 Metrics

Given the set of test users, their real BDI questionnaires and the automatic BDI questionnaires, the following effectiveness measures were calculated:

Average Hit Rate (AHR): Hit Rate (HR) is a rigorous metric that calculates the proportion of the 21 instances in which the automated questionnaire provides identical answers to those in the actual questionnaire. For instance, if an automated questionnaire yields 5 matches, the HR would be 5/21.

Average Closeness Rate (ACR): The Closeness Rate (CR) comes into play because the BDI responses represent an ordinal scale. If the actual user's response was "0" and a system responds with "3" then it should incur a more significant penalty compared to a system that responds with "1". For each question, the CR calculates the absolute difference (ad) between the actual and automated responses (e.g., $ad = 3$ for S1 and $ad = 1$ for S2), subsequently transforming this absolute difference into an effectiveness score using the formula: $CR = (mad - ad)/mad$. Here, mad represents the maximum absolute difference, the total count of potential answers minus one.

Average Difference in Overall Depression Levels (ADODL): While the preceding metrics evaluate the systems' capability to respond to each question in the BDI survey, the difference in overall depression level (DODL) takes a different approach. It does not focus on question-specific matches or disparities. Instead, it calculates the cumulative depression level (sum of all responses) for both the authentic and automated questionnaires. Next, it determines the absolute difference ($ad\_overall$) between the two depression scores. The overall depression score is between 0 and 63 and, thus, DODL is obtained as a normalized score in [0,1] as follows: $DODL = (63 - ad\_overall)/63$.

Depression Category Hit Rate (DCHR): In Psychology, it is standard practice to organize the overall depression scores into the following categories: Minimal Depression (depression levels 0-9), Mild Depression (depression levels 10-18), Moderate Depression (depression levels 19-29), and Severe Depression (depression levels 30-63). The final metric of effectiveness involves calculating the proportion of test users where the automated questionnaire assigned a depression category that matched the category determined by the actual questionnaire. These four metrics were the official metrics in the eRisk task described above and, thus, we adopted them to validate our summarization-based solution.

4.3 Alternative estimates

As alternative estimates of the level of severity of each depression symptom, we adopted the following strategies (all variants, including the human expert, received each BDI summary and the target question as input):

Human Expert: As argued above, we compare the model's predictions with an expert's prediction that reads the same sequence of BDI-biased summaries. The expert is a psychologist who was presented with the summaries and was asked to fill in the response to each BDI item. This alternative estimation helps to measure how similar the answers of the system and the human (e.g., using Cohen's kappa score).

T5: It is a well-known model that incorporates an encoder and a decoder, and it was pre-trained on a diverse set of data, including both unsupervised and supervised tasks (Raffel et al., 2020). Each task was transformed into a text-to-text format to fit with the model's structure. This model was also fine-tuned on QASC for question answering (via sentence composition) downstream tasks.

BERT-SQuAD: BERT large model[4] that was fine-tuned on the SQuAD dataset (Rajpurkar et al., 2016) for question answering.

5 Evaluation

Table 1 shows the results of our approach and all baseline methods over the three datasets. It includes two variants of chatGPT (versions 3.5 & 4), the alternative automatic methods (BERT-SQuAD and T5), and the expert's evaluation. All variants used the same summaries to respond to the BDI questions. Note that all metrics range in $[0, 1]$ and the higher the better.

One noteworthy observation is that both variants of ChatGPT obtained better results than the other automatic models when it came to fine-grained metrics that compute the effectiveness over individual questions (AHR, ACR). These results highlight how close the answers given by these two models are to the answers provided by the users. The chatGPT models tend to yield high values in the ACR metric. This is an important outcome since ACR focuses on the closeness between the real and automated responses, and a system with high ACR might have some potential to understand the feelings of the individual about the BDI symptom and develop psychological screening tools accordingly. On the other hand, BERT-SQuAD excelled in terms of global metrics (ADODL, DCHR) that focus on the divergence between the overall estimates of depression. It's noteworthy that ChatGPT version 3.5 consistently excels in producing responses that closely align with user input, while, version 4 tends to perform better when evaluated using broader global metrics, possibly owing to its enhanced capacity for generalizing information.

Still, there is much room for improvement in accurately predicting human responses. This is partly due to limited data availability, as there are many BDI topics that are not discussed or disclosed on social media. In any case, it is important to note that some automatic systems were on par with (or superior to) the assessment is done by human experts. In fact, the best automatic systems yielded equivalent performance to the expert psychologist in the fine-grained metrics (AHR and ACR) and better performance in the overall depression estimates (ADODL and DCHR).

In any case, the overall predictions (as reflected by DHCR) do not match those of the real surveys and this suggests that some BDI symptoms are difficult to grasp.

Regarding the time required, the expert took approximately 30 to 42 hours for each dataset (approximately 35 minutes per user). Instead, the LLMs took around 2-3 minutes to answer each user's questions. This signifies a substantial reduction in time, showcasing a pivotal advantage of automated methods that can facilitate the screening processes. By optimizing the extraction and analysis through computational tools, health professionals have the opportunity to allocate their saved time to the most confusing cases or just to review

[4] https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad

Models	AHR	ACR	ADODL	DCHR
eRisk 2019				
T5	0.2619	0.6198	0.7643	0.1000
BERT-SQuAD	0.2714	0.5963	**0.7740**	**0.2833**
chatGPT-3.5	**0.3857**	**0.6675**	0.7278	0.2000
chatGPT-4	0.3404	0.6556	0.7635	0.2000
expert	0.3833	0.6603	0.7270	0.2000
participants (mean)	0.3345	0.6416	0.7454	0.2611
participants (best)	**0.4143**	**0.7127**	**0.8103**	**0.4500**
eRisk 2020				
T5	0.3211	0.6578	0.7857	0.2143
BERT-SQuAD	0.3210	0.6325	**0.7947**	**0.2714**
chatGPT-3.5	**0.3748**	**0.6766**	0.7315	0.1857
chatGPT-4	0.3571	0.6728	0.7934	0.2143
expert	0.3694	0.6667	0.7082	0.1571
participants (mean)	0.3432	0.6688	0.7963	0.2807
participants (best)	**0.3830**	**0.6941**	**0.8315**	**0.3571**
eRisk 2021				
T5	0.2369	0.6008	**0.7377**	**0.2125**
BERT-SQuAD	0.2155	0.5605	0.7351	0.2000
chatGPT-3.5	**0.2714**	**0.6137**	0.6704	0.1375
chatGPT-4	0.2649	0.6014	0.7117	0.1125
expert	0.2500	0.5851	0.6161	0.075
participants (mean)	0.3107	0.6555	0.7586	0.2196
participants (best)	**0.3536**	**0.7317**	**0.8359**	**0.4125**

Table 1: Effectiveness results for the three datasets and comparison with the participants in the eRisk shared-task. We bold the best result of our models and participants of each year for an easier comparison.

the output of the LLMs.

Last, we have done an additional comparison between the predictions generated by the chatGPT 3.5 model and those of the domain expert. This comparison allows us to understand the degree of similarity between the respective responses. To that end, we employed Cohen's kappa score. The purpose is to provide insights into the model's performance by examining its alignment with human expertise across the entire range of users. These scores consistently hover around 0.28 for the 2019 and 2020 datasets and 0.0648 for 2021. This value, although modest, signifies a fair level of agreement between our model's predictions and those of the expert. In the 2021 dataset, we observe a low level of agreement; however, it is noteworthy that even in this collection the automated systems consistently outperformed the experts in predicting symptoms. These agreement levels underscore the model's capability to generate responses that align with expert judgments, demonstrating its reliability and effectiveness in providing valuable insights. While the agreement is not high, the model's performance is promising, considering the inherent complexity of the task at hand. These findings reinforce the model's potential to assist decision-making in the mental health domain.

Comparison against eRisk participants: To put these results in perspective, Table 1 also presents a comprehensive comparison between the models and the participants in the shared tasks of severity estimation in the eRisk editions of 2019, 2020, and 2021. Overall, our model demonstrates a good level of performance, outperforming the average results obtained in 2 out of 3 datasets. However, the top-performing participants achieved higher scores. This indicates that there is potential for enhancing our models' capabilities further. It is essential to mention that the participants performed extensive feature engineering and worked from the entire thread of user publications. In our study, this luxury was not extended to the LLMs or the human. In fact, it would be infeasible to ask the expert psychologist to read the entire history of posts, which

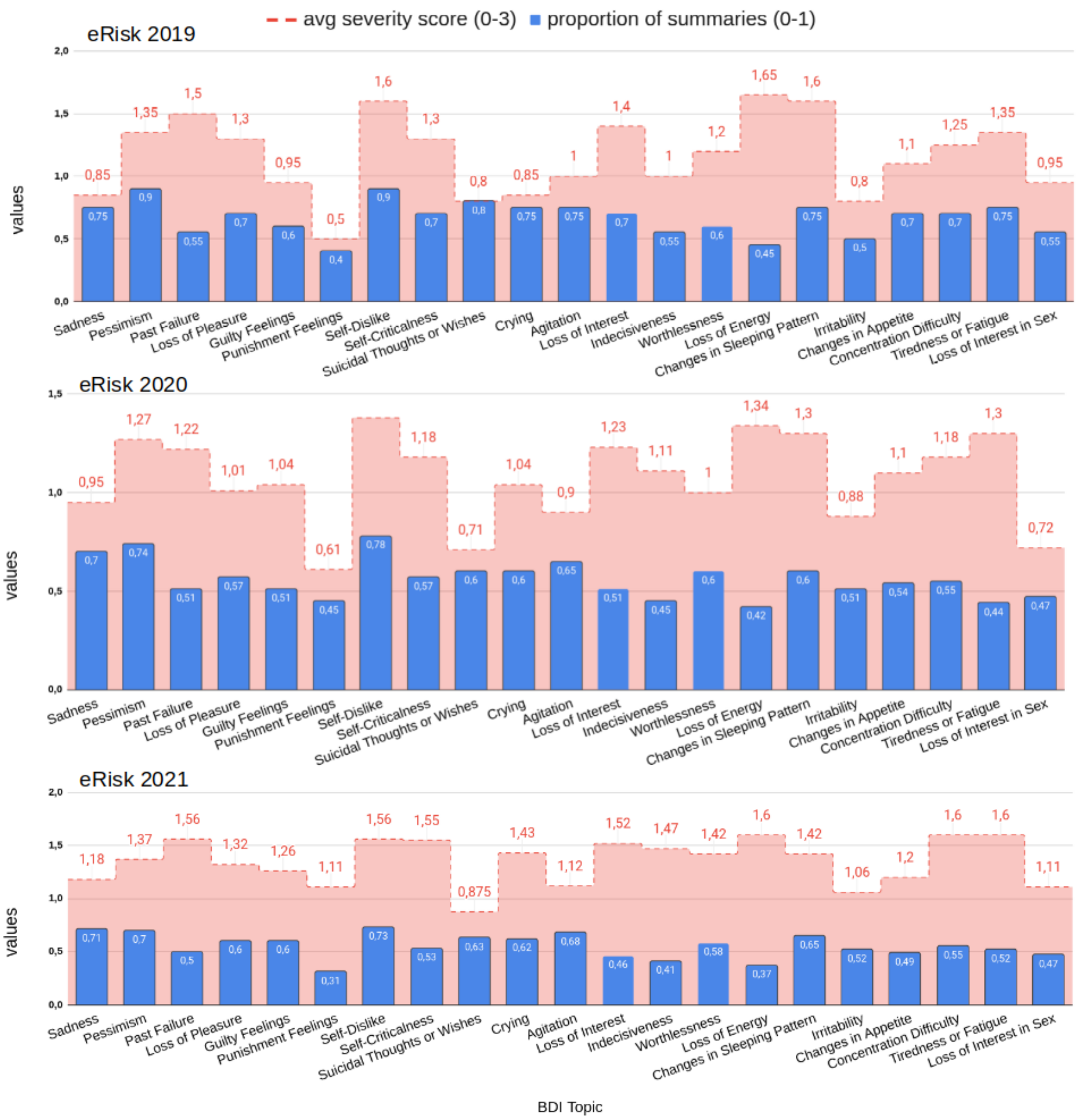


Figure 4: The X-axis represents the 21 BDI topics. The red bars show the average severity [0,3] of the corresponding symptom (as reflected in the ground truth) while the blue bars show the proportion of users [0,1] that had a non-empty summary for the symptom.

consists of thousands of publications. This human limitation motivates our work to employ advanced mining tools and implement search techniques that target adequate samples or key representative extracts, thus, summarizing the main themes within the users' history.

Nonetheless, this also opens up opportunities for refining our model's architecture and incorporating additional techniques to bridge the gap between its current performance and that of the most effective eRisk systems.

6 Analysis and Discussion

It is important to assess the extent to which the BDI topics have relevant sentences and the individual impact of BDI questions on the overall depression score. To that end, we analyze here the presence of relevant sentences for each BDI topic and plot it against the average rating in the ground truth (see Figure 4). The blue bars represent the proportion of users that had at least one relevant sentence for the corresponding topic (i.e. a non-empty summary). For instance, in 2019, for the topic of 'sadness', only 75% of the users had at least one related sentence. The red bars represent the average severity score provided by the users.

Certain themes, such as pessimism and self-dislike, are prominent (consistently provide relevant sentences for the majority of users) and tend to receive higher severity scores compared to other

BDI symptoms. This suggests a correspondence between the real feelings of these users and their social media activity (i.e., they tend to disclose thoughts about these symptoms). Other topics, such as loss of energy or punishment, have fewer relevant sentences (less than half of the users have at least one relevant sentence for these topics). Interestingly, in the case of energy loss, users provided high severity estimates, but the model could not find much evidence. This highlights a significant barrier in screening depression symptoms. If the model cannot find pertinent information on these topics then it can hardly supply a reliable estimate. In those cases, we assumed a rating of 0 and, thus, the models might be underestimating the state of the individual. In the future, it will be interesting to study other alternatives, such as estimating the overall depression scores based on partially filled questionnaires or estimating the missing BDI symptoms based on the most similar symptoms.

7 Conclusions and Future Work

In this study, we address a critical global concern, the prevalence of depression. We are committed to inducing a positive impact on automated methods for depression screening. To that end, we need a deeper understanding of depression symptoms and more evidence of how the symptoms reflect on social media. We have presented a comprehensive approach that involves extracting BDI-biased summaries from users' publications and exploiting different large language models to estimate the responses of those users to the Beck Depression Inventory. Our evaluation across various depression datasets yielded promising results, showcasing our method's potential to contribute to the understanding and assessment of depression. Some of the proposed variants compete favorably with state-of-the-art methods and expert human evaluations. This work represents a valuable step forward in leveraging the power of data to address mental health challenges on a broader scale. In future work, we want to explore the application of other lexical resources that are even more specialized for the task of extraction of relevant sentences, as well as the usage of clinical data to train more specialized language models.

Furthermore, the primary focus of our work revolves around leveraging these summaries. Specifically, our interest lies in the potential application of this tool to extract valuable linguistic indicators. This application could be useful in enhancing psychologists' understanding of how depression manifests in social media contexts. By delving into linguistic patterns and cues from user-generated content we could offer valuable insights that contribute to the refinement of psychologists' working knowledge. We also are interested in expanding this study to different languages, since most of the work related to mental disorders has focused on English.

Finally, this study represents a preliminary exploration but we believe that the ability to model user behavior through social media analysis offers promising prospects for the development of future wellness-oriented technologies. This innovative technology has the potential to function as a preemptive warning system, conducting extensive analyses and delivering pertinent information concerning mental health without compromising user privacy. For example, we could design local, regional, or national estimates of the prevalence of multiple depression symptoms, allowing authorities to make informed decisions about professional assistance, emotional support campaigns, and so forth. Under this context, users should always retain autonomy in choosing to have access to certain recommendations or preemptive measures, empowering them to make informed decisions about their well-being.

Ethic Statement and Impact

Examining social media content raises potential privacy and ethical concerns. This research is exempt from IRB review because we only experimented with existing publicly available collections and did not contact any social media users. The datasets only contain public user interactions and we have diligently adhered to the terms of use and user agreements of these collections. Moreover, these collections are anonymized. While public posts may be freely available to anyone, individuals may not intend for them to have a broad audience. We have therefore paraphrased the extracts shown in this paper. With this research, we also want to make a positive impact on society, and one significant contribution we may provide is to better understand depression. Specifically, we want to learn information that will aid mental health diagnosis and help those challenged by mental illness.

Limitations

It is essential to acknowledge certain constraints inherent to this study. Notably, the research is observational, lacking access to personal and psychological data typically incorporated in risk assessment investigations. Furthermore, an unavoidable bias stems from the data source (only users who are exposed to social media and, specifically, to Reddit were included in the study). Segments of the population, such as elderly people or individuals who consciously abstain from maintaining online accounts or opt to keep their profiles private, cannot be monitored.

Acknowledgements

The second author thanks the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G/01, ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center in ICT of the University of A Coruña as a Research Center of the Galician University System and the project PID2022-137061OB-C21 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund).

The first and third authors thank: i) the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G-2019/04, ED431C 2022/19) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System, and ii) the financial support supplied by project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund).

The third author thanks the financial support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego).

The authors also thank the funding of project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU).

References

American Psychiatric Association APA. 2020. What is depression?

Aaron Beck, C.H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An inventory for measuring depression. *Arch Gen Psychiatry*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, USA*.

Fabio Crestani, David E. Losada, and Javier Parapar. 2022a. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project*. Springer Verlag, Englewood Cliffs, NJ.

Fabio Crestani, David E Losada, and Javier Parapar. 2022b. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*, volume 1018. Springer Nature.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Ronald C Kessler, Evelyn J Bromet, Victoria Shahly Peter de Jonge, and Marsha. 2017. The burden of depressive illness. *Public Health Perspectives on Depressive Disorders*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 340–357, Cham. Springer International Publishing.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 272–287, Cham. Springer International Publishing.

Colin D Mathers and Dejan Loncar. 2006. Projections of global mortality and burden of disease from 2002 to 2030. *PLOS Medicine, Public Library of Science*, pages 1–20.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.

Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2018. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 647–656, New York, NY, USA. Association for Computing Machinery.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2021. Overview of erisk 2021: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 324–344, Cham. Springer International Publishing.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 294–315, Cham. Springer Nature Switzerland.

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

Anxo Pérez, Neha Warikoo, Kexin Wang, Javier Parapar, and Iryna Gurevych. 2022b. Semantic similarity models for depression severity estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16104–16118.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Miguel Enrique Renteria-Rodriguez. 2018. Salud mental en mexico. *NOTA-INCyTU NÚMERO 007*.

Esteban A. Ríssola, David E. Losada, and Fabio Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2).

Ruba Skaik and Diana Inkpen. 2020. Using social media for mental health surveillance: A review. *ACM Comput. Surv.*, 53(6).

Sumithra Velupillai, Gergö Hadlaczky, Genevieve M. Gorrell, Nomi Werbeloff, Dong Nguyen, Rashmi Patel, Daniel Leightley, Johnny Downs, Matthew Hotopf, and Rina Dutta. 2019. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. *Frontiers in Psychiatry*, 10.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Ke Zhu. 2022. Psychiatric scale guided risky post screening for early detection of depression. In *International Joint Conference on Artificial Intelligence*.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *CoRR*, abs/2101.00774.