DepressMind: A Depression Surveillance System for Social Media Analysis

Roque Fernández-Iglesias, Marcos Fernández-Pichel,
Mario Ezra Aragón and **David E. Losada**
Centro Singular de Investigación en Tecnoloxias Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
roque.fernandez.iglesias@rai.usc.es,
{marcosfernandez.pichel,ezra.aragon,david.losada}@usc.es

Abstract

Depression is a pressing global issue that impacts millions of individuals worldwide. This prevailing psychological disorder profoundly influences the thoughts and behavior of those who suffer from it. We have developed DepressMind[1], a versatile screening tool designed to facilitate the analysis of social network data. This automated tool explores multiple psychological dimensions associated with clinical depression and estimates the extent to which these symptoms manifest in language use. Our project comprises two distinct components: one for data extraction and another one for analysis. The data extraction phase is dedicated to harvesting texts and the associated meta-information from social networks and transforming them into a user-friendly format that serves various analytical purposes. For the analysis, the main objective is to conduct an in-depth inspection of the user publications and establish connections between the posted contents and dimensions or traits defined by well-established clinical instruments. Specifically, we aim to associate extracts authored by individuals with symptoms or dimensions of the Beck Depression Inventory (BDI).

1 Introduction

Psychological analysis is a multifaceted discipline dedicated to the study of human behavior and the intricacies of related mental processes. This field holds immense significance as it helps to gain deeper insights into individuals and the functioning of their minds. Such insights are instrumental in both the treatment and prevention of mental health issues. Among the most pressing concerns within psychological analysis is depression, a pervasive mood disorder that affects countless individuals worldwide (World Health Organization, 2023). Depression constitutes a serious affliction that significantly diminishes a person's quality of life and can have profound consequences on his physical and mental well-being. Although there is a wide range of available therapeutic approaches, ranging from psychotherapy to pharmaceutical interventions, the burden of inadequately treated or unrecognized depression is still substantial (Evans-Lacko et al., 2018). Recent years have witnessed a noticeable increase in the global prevalence of depression (World Health Organization, 2022), reflecting the need for further research and design of new screening tools.

The Internet and Social Media (SM) have sparked a revolutionary transformation in the way we communicate. Online sources have also opened up new possibilities in psychological analysis, particularly for extracting behavioral patterns from specific demographic groups. One of the primary advantages is the availability of a large amount of data and multiple pieces of evidence about human behavior. These valuable resources can be exploited to perform observational studies that support mental health research. SM platforms are spaces where individuals can freely express themselves and publicly share their thoughts, feelings, and opinions. Some popular sites and forums have become meeting points for disclosing personal preoccupations. Notably, online communities focused on topics like depression and other mental health issues have emerged, enabling individuals to share their personal experiences and receive support from others who have navigated similar challenges (Skaik and Inkpen, 2020; Ríssola et al., 2021; Crestani et al., 2022).

It is crucial to develop new screening mechanisms to early detect the increasing severity of depressive symptoms and provide explanatory extracts of the evidence found. To that end, we exploit advanced Natural Language Processing (NLP) solutions to analyze patterns and trends of depression. Such automated screening would allow to moni-

[1]A demonstration video of DepressMind is available at: https://youtu.be/rd6ZVWbxYrM

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
System Demonstrations, pages 35–43
March 17-22, 2024

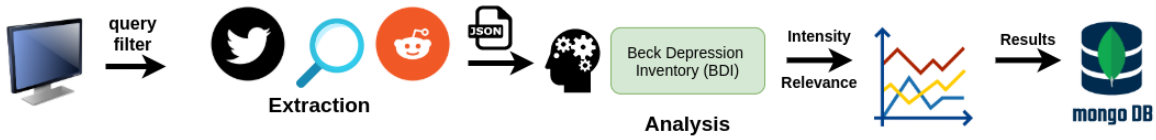

Figure 1: General diagram of DepressMind's architecture. First, it collects data from different sources (social networks), and then, it analyzes the levels of intensity and relevance in depression according to the BDI test.

tor a large number of people and identify those who may be at risk of developing a mental disorder. In practice, this could lead to interventions that provide individuals with the right support in time. Alternatively, the insights gained by our tool could inform preventive measures and help detect risk factors. For example, by understanding the evolution of depressive symptoms in specific segments of the population (e.g. individuals of a certain age) we could instigate the definition of new communication or mitigation measures by relevant health authorities. In any case, this exploitation would have to be done with proper ethical and privacy precautions. The goal of this paper is to exemplify the potential contributions of this technology, while its actual deployment is beyond the scope of the present project. Furthermore, DepressMind is available for research purposes and, thus, those interested in extracting and analyzing online data from a BDI perspective can get access to this new screening tool[2].

DepressMind stands on a retrieval method that supports a semantic representation for each category of the BDI-II questionnaire. BDI (Beck et al., 1961) is a recognized clinical instrument designed to assess the manifestation of 21 depressive symptoms, such as sadness, pessimism, or loss of energy. Given a textual excerpt, DepressMind computes two different metrics, relevance, and intensity, with respect to each BDI symptom. The whole process is depicted in Figure 1. We can summarize our contributions as follows:

- Development of a Reddit and Twitter extraction module designed to compile a comprehensive collection of online publications using specific filtering criteria (e.g., keywords, user accounts, communities, etc.).

- Implementation of an advanced analytical tool capable of assessing psychological dimensions and their temporal evolution. This analysis leverages semantic similarity estimates and linguistic models to provide valuable insights into user-generated content.

- Integration of the above components into a user-friendly web application that streamlines the process of analyzing Beck Depression Inventory (BDI) dimensions over a precompiled dataset of social media posts. This application allows users to visualize results through graphs and statistical representations for a more comprehensive understanding of the data.

[2]https://github.com/roque-fernandez/DepressMind

2 Related Work

In recent times, the field of digital mental health analysis has experienced significant expansion. Automated NLP techniques have emerged as effective means for identifying traces of mental health disorders from user-generated online publications. For instance, in a pioneering study, Choudhury et al. (2013) conducted initial research on automatically detecting depression. These researchers collected reliable data on depression through crowd-sourcing, employing the CES-D (Center for Epidemiological Studies Depression Scale) inventory to assess depression levels (Radloff, 1977).

Gaur et al. (2018) introduced an unsupervised method aimed at aligning the content from different mental health-related subreddits with the most suitable DSM-5 (Diagnostic and Statistical Manual of Mental Disorders - 5th Edition) categories (Association, 2013). Given the knowledge encoded into the DSM-5 manual and other meticulously curated medical resources, a specialized lexicon was constructed. This language resource included n-grams associated with each mental health disorder listed in the DSM-5. Additionally, the authors enhanced an existing ontology related to drug abuse, incorporating terminology and slang expressions derived from Reddit. The resulting lexicon was exploited to quantify the connection between the content of the subreddits and the DSM-5 categories.

In a study conducted by Perez et al. (Pérez et al., 2022), an innovative approach was introduced to automatically measure the severity of depression among social media users. This research team addressed the challenge of quantifying the intensity of depression indicators and explored the application of neural language models to capture various aspects of the user-generated content. This team presented two alternative methods for evaluating depression symptoms, based on the individual's willingness to openly discuss them. The first method relied on analyzing global language patterns present in the user's posts, while the second method focused on identifying direct mentions of symptom-related concerns. Both approaches resulted in automatic estimates of the overall BDI-II score.

In a related vein, Pérez et al. (2023) introduced a platform designed to assist in the assessment and monitoring of depression symptoms among social media users. The tool aimed at capturing a broad context of the individuals by incorporating user profiling capabilities. The authors claimed that the platform could assist professionals in labeling data and utilizing depression estimators and profiling models.

These initiatives have piqued our interest in exploring the potential of semantic matching between relevant medical questionnaires and user-generated publications. Our system, DepressMind, allows us to extract and visualize BDI-related extracts and we present here a preliminary evaluation that shows its potential to automatically estimate an individual's response to the BDI-II items.

3 Design and Implementation

3.1 Data collection

To facilitate data collection, we developed a dedicated module that allows us to gather information from Twitter and Reddit. This module serves as an intermediary, streamlining the process of parsing user-generated content. The tracking process adapts based on the data source. In the case of Twitter, our module utilizes the free access Twint library[3] to connect with Twitter's API and retrieve a specified number of tweets based on user-defined filters. The collected data is then stored in a JSON file. For Reddit, we have implemented a custom tracker designed specifically for this platform. This tracker retrieves Reddit posts, applies the necessary filters, and stores results in JSON format. Furthermore, our system produces a separate file designed for recording statistics for the resulting collection.

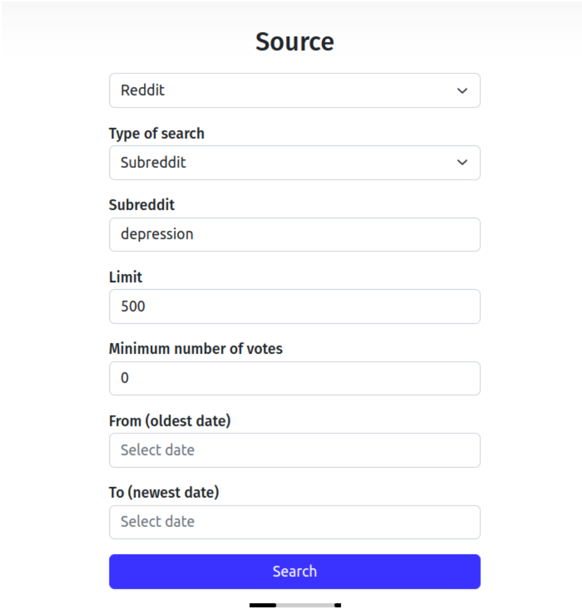

Figure 2: DepressMind's extraction interface.

The Reddit crawler parses the HTML code and extracts class patterns and tags associated with Reddit posts. To that end, we employ BeautifulSoup, a Python library that simplifies the parsing and data extraction from XML and HTML documents. Given the evolving nature of Reddit, consistent data extraction has become challenging. Our extraction works from the OldReddit interface. In May 2018, Reddit introduced a new interface and appearance that met with significant disapproval (over 80% of users). Consequently, the site decided to maintain the old interface on a separate domain. To this day, approximately 15% of users continue to access Reddit via the old interface, with no plans to discontinue. This legacy interface provides us with a unique advantage, as its HTML structure is more amenable to automated crawling.

DepressMind supports four extraction possibilities: by subreddit, by user account, by keywords, and a general (unrestricted) extraction. These extraction options share certain parameters, such as the maximum number of posts to be retrieved, and the range of dates of the extracted publications. Figure 2 shows an instance of the subreddit-based

[3] At the time of building DepressMind, Twint allowed free access to publications and downloads. Due to changes in Twitter's policies (now referred to as "X"), this module is temporarily disabled. We are actively monitoring for future updates that may reinstate its functionality.

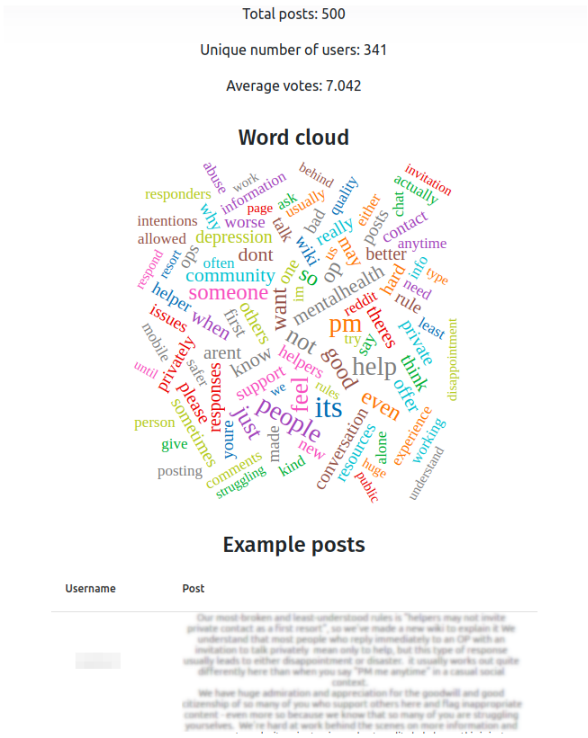

Figure 3: Example of the output after the extraction (the user posts and accounts have been pixelated for privacy reasons).

option. After the extraction, DepressMind displays the number of posts found, a word cloud, some example posts, and other general statistics (see Figure 3).

3.2 Analysis of Psychological Dimensions

The extraction functionalities described above allow any user to produce and download a JSON-formatted dataset and, thus, DepressMind can be employed by researchers or practitioners to build their collections. In any case, our primary goal was to make a depression screening tool and, thus, DepressMind includes a BDI-based analytical module. To that end, DepressMind assesses the semantic proximity between the texts gathered from social media and the textual content of the BDI questionnaire. Each BDI item contains a title (e.g. “Guilty Feelings”) and some possible responses (e.g. “0. I don’t feel particularly guilty”, “1. I feel guilty a good part of the time”, “2. I feel quite guilty most of the time”, or “3. I feel guilty all the time”). These texts are used to assess the relevance and intensity of each topic, as described below.

3.2.1 Relevance

This analysis aims to address the question: How relevant are individual depression symptoms within

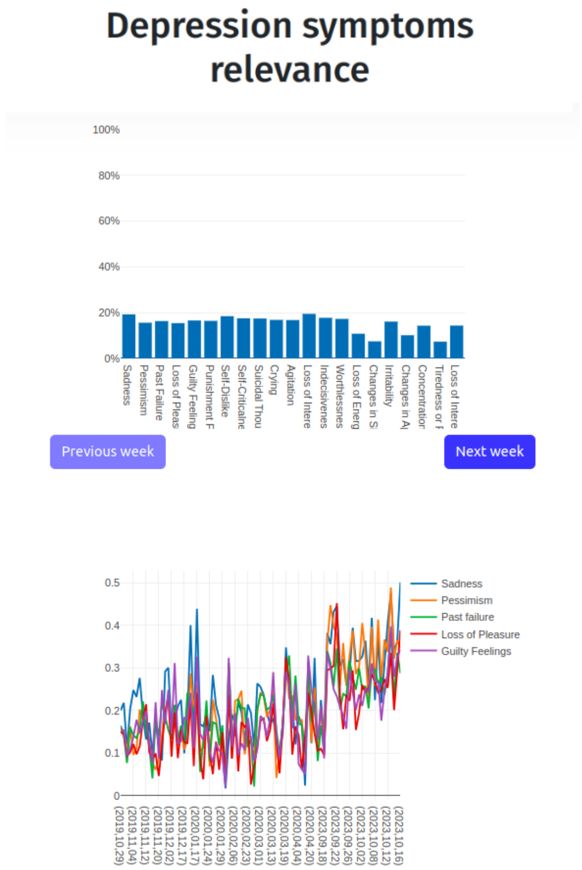

Figure 4: Relevance analysis and evolution over time of each BDI topic.

our collection? Given a collection of posts, DepressMind outputs an array of 21 values, one for each BDI symptom. Each element in the array estimates the topical presence of the BDI item in the corpus. These estimates are produced as follows. First, DepressMind transforms each question from the BDI into an embedding representation utilizing SentenceBERT (Reimers and Gurevych, 2019), an adaptation of the pre-trained BERT model designed to produce semantically meaningful sentence embeddings[4]. For every topic, we generate embedding representations of the title of the corresponding BDI item and each potential response. The similarity between a sentence in the corpus and a BDI item is computed as the maximum similarity (using cosine similarity) between the embedding of the sentence (e_s) and the embeddings of the BDI title and the BDI responses: $score_{rel}(s, BDI_n) = max\{sim(e_s, e_{title_n}), sim(e_s, e_{response1_n}), ...\}$.

The relevance of a BDI item in the corpus (C) is computed as the average relevance computed over the entire set of sentences: $rel(BDI_n, C) =$

[4] More specifically, with the all-MiniLM-L6-v2 model

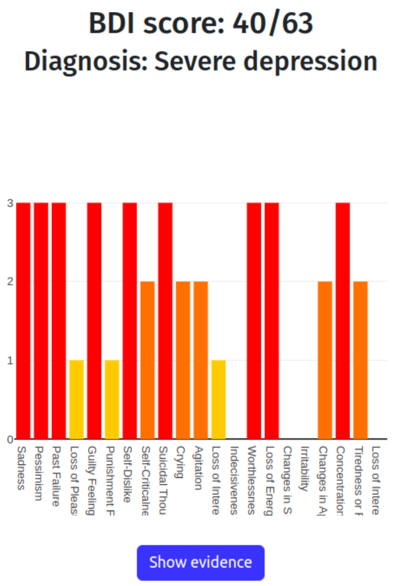

Figure 5: Intensity analysis.

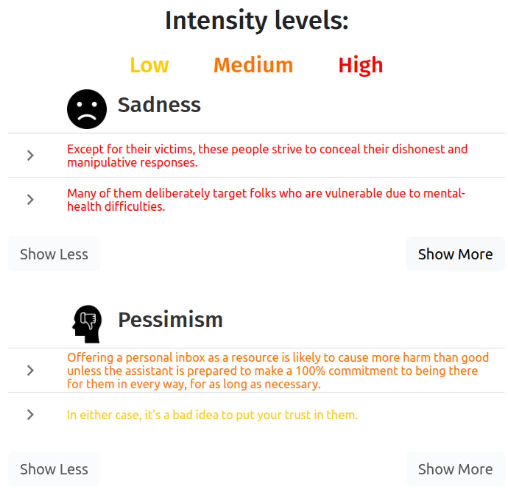

Figure 6: Top scoring sentences for two BDI items (sadness and pessimism). For privacy reasons, texts have been paraphrased.

$avg_{s \in C}\ score_{rel}(s, BDI_n)$. Figure 4 shows an example of this analysis.

3.2.2 Intensity

This metric estimates the intensity of each depression symptom in the input corpus. While the relevance of a symptom informs us about topicality (i.e., the presence of sentences related to the BDI item), intensity provides insight into the level of severity. Many relevant sentences might induce different intensity levels. For instance, "I never sleep" and "I have occasional sleeping problems" are both relevant to the BDI item on sleeping problems but the first one should be assigned a higher intensity score. Given a collection of posts, DepressMind outputs an array of 21 intensity values, one for each BDI symptom. These scores are integer values ranging from 0 to 3. This is the standard ordinal scale of the responses in the BDI questionnaire. These estimates are produced as follows.

For each BDI item, we only consider as candidates the sentences whose relevance $score_{rel}(s, BDI_n)$ is greater than a given threshold. For each candidate sentence, we try to align it with the possible responses of the BDI item. To that end, we employ an entailment model[5] with three possible output categories: *contradiction*, *entailment*, *neutral*. If the candidate sentence does not entail the response (output category is *contradiction* or *neutral*) then the alignment is discarded. Otherwise, we apply a softmax function to the entailment value, which returns the probability that the sentence s supports the response. DepressMind aligns the sentence with the response that has the highest probability of entailment and outputs the response's severity level, $val_{int}(s, BDI_n) \in \{0, 3\}$. For example, for the "Suicidal thoughts" symptom, a sentence such as "I want to kill myself" would be aligned with the response "I would kill myself if I had the chance" and, thus, it would be assigned an intensity value of 3. Finally, the BDI item's intensity score of the entire corpus is computed as the maximum intensity over the available sentences: $int(BDI_n, C) = max_{s \in C}\ val_{int}(s, BDI_n)$. The rationale is that a single severe sentence should be a matter of concern. Figure 5 shows an example of the intensity analysis. DepressMind also shows explanatory evidence, by presenting the sentences that led to the final intensity estimation (see Fig 6).

4 Empirical Evaluation

To validate the tool, we utilized the eRisk evaluation datasets (2019, 2020, and 2021) (Losada et al., 2019, 2020; Parapar et al., 2021). These evaluation tasks aim to estimate the severity of the 21 BDI symptoms for a set of social media users. For each user, the datasets include a self-report BDI questionnaire and the user's history of publications (participants gave their consent for tracking of their public posts). We fed each user's collection to DepressMind, obtained the estimated levels of severity, and compared them with those filled by the users. The next subsection describes the four effectiveness metrics. Further details about them

[5] We used the cross-encoder/nli-deberta-v3-base model.

method	AHR	ACR	ADODL	DHCR
		eRisk 2019		
ours	0.3143	0.6492	0.7476	0.2000
participants	0.3345	0.6416	0.7454	0.2611
		eRisk 2020		
ours	0.3156	0.6574	0.7481	0.2571
participants	0.3432	0.6688	0.7963	0.2807
		eRisk 2021		
ours	0.2601	0.6319	0.7383	0.3375
participants	0.3107	0.6555	0.7586	0.2196

Table 1: Effectiveness results over the three datasets. The lower result reports the mean performance achieved by the participants in the eRisk shared task.

are available at (Losada et al., 2019, 2020; Parapar et al., 2021).

4.1 Metrics

Average Hit Rate (AHR): AHR measures how often the automated questionnaire produces identical responses to the real questionnaire.
Average Closeness Rate (ACR): ACR takes into consideration the ordinal scale and penalizes more the errors where the system's response deviated significantly from the actual user's response.
Average Difference in Overall Depression Levels (ADODL): While the previous metrics focus on effectiveness at the BDI item level, ADODL calculates the total depression level, which is the sum of all responses, for both the real and automated questionnaires, and measures the difference.
Depression Category Hit Rate (DCHR): DCHR considers the four possible overall depression levels (0–13: minimal depression, 14–19: mild depression, 20–28: moderate depression, 29–63: severe depression) and measures the proportion of users for whom DepressMind assigned a depression category that matches the category determined by the real questionnaire.

4.2 Results and Analysis

Table 1 shows the results of DepressMind and compares them with the mean results of the participants in the three evaluation campaigns. In general, our approach yields competitive results in comparison with the systems implemented by the participating teams. Furthermore, the participants' systems employed extensive feature engineering leading to black-box models, while DepressMind visually explains its decisions and presents evidence that can be validated by an expert. We leave the exploration of more complex solutions (e.g., other sentence embedding models) for future work. Regardless, there exists potential for enhancing the structure of our model and integrating further methods to narrow the performance gap between the current version of DepressMind and the most effective eRisk systems. In any case, the challenge is intrinsically difficult, as most users do not publicly disclose information about many BDI topics. From our perspective, the most compelling feature of this screening technology is not its capacity for automatic diagnosis, but rather its proficiency in highlighting the most significant concerns, discerning temporal trends, and providing valuable insights to expert users (e.g., to psychologists).

To further analyze these collections, let us focus on the individual influence of specific BDI items. To this end, Figure 7 plots the percentage of users with relevant evidence for each BDI item (blue bars) and the percentage of intense sentences for each BDI item in the collection (orange bars). This helps to shed light on the BDI items that are more prominently discussed on social media. Topics such as feelings of worthlessness, sadness, agitation, and punishment are the most recurrent and also produce most of the severe sentences. In contrast, topics like crying, self-criticism, and loss of interest in sexual activities have little presence and intensity. Notably, in all the datasets, most topics have less than half of the users with relevant evidence. This underscores how challenging it is to infer depression symptoms from social media evidence. With no relevant information on these topics, a model can hardly provide a reliable assessment. In such cases, DepressMind assumes a score of 0, potentially leading to an underestimation of the individual's state. In the future, it would be worthwhile to explore alternative imputation approaches, such as estimating overall depression scores based on partially completed questionnaires or estimating missing BDI symptoms based on related topics.

5 Conclusion

In this study, we try to contribute to the understanding of a pressing global issue: the prevalence of depression. Our main goal was to make a positive contribution in the area of automated methods for screening depression. To that end, we introduced DepressMind, an analytical tool that extracts social media posts and assesses their relevance and severity concerning the 21 standardized BDI symptoms. An evaluation of several depression datasets has

2019			2020			2021		
% entailments	Dimension	% users with evidence	% entailments	Dimension	% users with evidence	% entailments	Dimension	% users with evidence
0.59	Worthlessness	95.00	0.26	Worthlessness	81.43	0.37	Worthlessness	75
0.27	Sadness	75.00	0.16	Punishment Feelings	80.00	0.48	Agitation	73.75
0.16	Punishment Feelings	70.00	0.22	Agitation	65.71	0.20	Punishment Feelings	68.75
0.30	Agitation	70.00	0.16	Sadness	60.00	0.21	Sadness	65
0.16	Pessimism	65.00	0.12	Self-Dislike	45.71	0.26	Self-Dislike	56.25
0.12	Guilty Feelings	60.00	0.08	Changes in Sleeping Pattern	45.71	0.12	Irritability	53.75
0.07	Suicidal Thoughts or Wishes	55.00	0.05	Loss of Pleasure	42.86	0.15	Loss of Energy	51.25
0.08	Changes in Sleeping Pattern	55.00	0.05	Guilty Feelings	42.86	0.10	Changes in Sleeping Pattern	48.75
0.29	Self-Dislike	50.00	0.05	Irritability	42.86	0.08	Loss of Pleasure	45
0.08	Loss of Energy	50.00	0.06	Changes in Appetite	42.86	0.06	Changes in Appetite	45
0.03	Loss of Pleasure	40.00	0.04	Tiredness	42.86	0.15	Concentration Difficulty	45
0.03	Loss of Interest	40.00	0.04	Pessimism	40.00	0.08	Tiredness	45
0.07	Irritability	40.00	0.03	Past Failure	40.00	0.09	Pessimism	43.75
0.07	Past Failure	35.00	0.03	Suicidal Thoughts or Wishes	40.00	0.04	Past Failure	41.25
0.04	Concentration Difficulty	35.00	0.04	Loss of Energy	38.57	0.10	Suicidal Thoughts or Wishes	41.25
0.07	Changes in Appetite	25.00	0.03	Loss of Interest	37.14	0.06	Guilty Feelings	38.75
0.03	Tiredness	25.00	0.05	Concentration Difficulty	37.14	0.05	Self-Criticalness	37.5
0.03	Crying	20.00	0.02	Loss of Interest in Sex	30.00	0.02	Loss of Interest in Sex	36.25
0.01	Indecisiveness	20.00	0.02	Self-Criticalness	25.71	0.08	Loss of Interest	33.75
0.01	Self-Criticalness	15.00	0.01	Indecisiveness	21.43	0.02	Indecisiveness	25
0.01	Loss of Interest in Sex	10.00	0.02	Crying	20.00	0.02	Crying	23.75

Figure 7: Intensity and Relevance of the 21 BDI items.

produced promising results, highlighting the potential to enhance our understanding, assessment, and management of various depression symptoms. This work represents an initial advancement towards harnessing the power of data to address mental health challenges on a broader scale. In our future work, we aim to explore the use of specialized lexical resources for more effective sentence extraction and to exploit clinical data for training more specialized language models. We are also interested in expanding this study to different languages, as most of the research studies on mental disorders have been centered on English.

Ethics Statement and Limitations

This form of analysis aims to support the development of emerging technologies designed to alert about early signs of depression and provide substantial supporting evidence. As highlighted by Neuman et al. (2012), it is important to view these novel screening methods not as "magic replacements for human experts", but rather as computational tools that can significantly alleviate the workload on public health systems, especially in terms of facilitating regional preventive measures. This project did not involve any interaction with social media users, such as interventions to provide health-related recommendations. Instead, it is an observational study focused on publicly accessible data and available research collections. The three evaluation datasets are publicly available and transparent, and our research strictly adhered to ethical guidelines, specifically those related to AI ethics by design. For instance, the data collections are anonymized. Furthermore, we put in place stringent measures to safeguard sensitive information (e.g., the example sentences reported in the paper were paraphrased). Our ultimate goal is to make a positive contribution to society, with a key emphasis on improving our understanding of depression.

Another limitation is that at the time of building the system, Twint allowed free access to Twitter publications and downloads. Due to changes in Twitter's policies (now referred to as "X"), the Twitter feature is temporarily inactive. We are continuously observing for potential revisions that could restore its operation.

Additionally, it is essential to acknowledge that the datasets retrieved using DepressMind or those used to evaluate the tool may exhibit inherent biases typical of social media data. For example, gender, age, or sexual orientation biases. Any researcher or practitioner using DepressMind to collect and analyze data should bear this in mind and apply proper fairness and bias correction measures.

Acknowledgments

The 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 fourth 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. 2013. *Diagnostic and statistical manual of mental disorders*, 5th edition. American Psychiatric Publishing, Washington.

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. 2022. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*, volume 1018. Springer Nature.

Sara Evans-Lacko, Sergio Aguilar-Gaxiola, Ali Al-Hamzawi, Jordi Alonso, Corina Benjet, Ronny Bruffaerts, Wai-Tat Chiu, and Silvia Florescu. 2018. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the who world mental health (wmh) surveys. *Psychological Medicine*, 48:1560–1571.

Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. "let me tell you about your mental health!": Contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 753–762.

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.

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

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.

Anxo Pérez, Marcos Fernández-Pichel, Javier Parapar, and David E. Losada. 2023. Depresym: A depression symptom annotated corpus and the role of llms as assessors of psychological markers. *ArXiv*, abs/2308.10758.

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

Lenore Sawyer Radloff. 1977. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401.

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.

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).

WHO World Health Organization. 2022. Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide.

WHO World Health Organization. 2023. Depressive disorder (depression).