Journal of Computational Social Science (2025) 8:48
https://doi.org/10.1007/s42001-025-00377-9

RESEARCH ARTICLE

On the incidence of depression symptoms on social media

Esteban A. Ríssola[1] · **Mario Ezra Aragón**[2] · **David E. Losada**[2] · **Fabio Crestani**[3]

Received: 24 July 2024 / Accepted: 11 February 2025 / Published online: 8 March 2025

Abstract
Due to their increasing popularity, researchers and health professionals are actively utilizing social media networks as valuable tools to recognize linguistic patterns associated with mental health. In this research, our aim was to better understand to what extent the Beck Depression Inventory (BDI) could undergo automated screening based on users' social media feeds. To this end, we conducted different experiments to analyze the prevalence of BDI items on social media. We present an approach to categorizing and ranking BDI items considering the quantity of information that can be obtained from social media posts. Given publications written by people who have personally reported being diagnosed with depression, we run different search methods and, based on the number of elements retrieved, we study the prevalence of BDI symptoms at two levels of coverage. Finally, we investigate the impact of prevalence and various characteristics on the efficacy of automated assessment tools. Our analysis indicates that specific elements occur consistently across various search methods and social media platforms, implying a higher prevalence of related symptoms in the data sets analyzed. Interestingly, some items with low incidence in the data sets are those of the BDI questionnaire, whose responses are more accurately estimated using automated methods.

Keywords Health informatics · Information retrieval · Social media mining · Depression · Beck Depression Inventory

1 Introduction

The exploitation of online social media platforms is altering the landscape of mental state evaluation methodologies. Researchers in mental and behavioral health can now study human behavior on a scale hardly imaginable years ago [1–3]. Throughout the day, individuals use social media platforms to discuss their current social conditions [4], to share their interests [5] or personal daily life experiences [6], and

Extended author information available on the last page of the article

even reveal their moods and feelings [7]. This provides a unique opportunity to model the dynamics of mental health [8] proactively.

New technologies focused on risk assessment and decision-making based on user-generated content promise to create a substantial impact. They can provide affordable and non-intrusive mechanisms for the widespread early screening of mental health [9, 10] or measure the presence, intensity, or patient satisfaction about treatments received [11]. For example, novel language technology solutions [12] have begun to be considered, at least to support preliminary screening processes or to promote understanding and acknowledgment of mental health concerns. In this context, studies in psycho-linguistics have suggested that understanding how individuals express themselves verbally serves as a means to assess various behavioral aspects [13]. Linguistic characteristics offer a distinct insight into thoughts and emotions, facilitating a direct evaluation of changes in mental states [14, 15].

Recently, the core principles of web search and data mining have been gradually extended. This has led to new applications that were difficult to envision a few years ago. For example, retrieval and extraction components have been instrumental in developing new and innovative applications related to web-based propaganda [16], personality analysis [17], or psychological aspects of information access [18, 19]. Initiatives such as the Early Risk Prediction on the Internet (eRisk) CLEF lab [20–26] and the Computational Linguistics and Clinical Psychology Workshop (CLPsych) [27] have played a key role in encouraging global collaboration to leverage data from social media and create models that estimate the prevalence of multiple indicators associated with mental disorders.

These efforts to develop online screening tools have mainly focused on extracting and encoding different attributes from users' social media feeds (e.g., to identify traces of mental disorders [28]). In 2019, eRisk initiated a unique project to investigate the potential for automated assessment of the intensity of various symptoms linked to depression. These symptoms were standardized clinical indicators from a well-known psychometric test (BDI) [29].

The BDI is part of a category of structured psychological assessments, commonly referred to as questionnaires or inventories. These tools offer a standardized and unbiased evaluation of a subset of human behavior [30]. These psychometric tests are widely employed for obtaining high-quality data from various channels, including online platforms [31, 32]. The BDI essentially functions as a self-assessment tool designed to measure typical attitudes and symptoms associated with depression. Using a series of 21 multiple-choice questions, it assesses the degree and prevalence of emotions such as sadness, hopelessness, self-disapproval, social withdrawal, and diminished energy. The BDI is based on the concept that negative cognitive distortions play a central role in depression [29].[1] It was created by collecting and organizing patients' accounts of their symptoms and then using this information to establish a scale capable of indicating the intensity of a specific symptom [33].

The eRisk shared-data tasks produced an initial result indicating the viability of automatically gauging the severity of certain symptoms, specifically those measured

[1] The complete questionnaire is available at https://bit.ly/3drfpVg.

by BDI items, through analyzing user interactions on social media. However, the performance achieved by the participating teams remains limited, and a highly effective tool for screening depression is yet to be developed. Motivated by this, we consider it important to gain new insight into the incidence of specific depression symptoms reflected on social media. It is crucial to perform new experiments and analyses of existing data sets to better understand automated screening methods, particularly to shed light on why they fail. Is it due to insufficient evidence regarding specific BDI items? Is it a result of systems misinterpreting the available evidence? By addressing these queries, we can steer future research in this field and propose paths for attaining more accurate assessments of the severity of various symptoms associated with depression. For instance, depending on the *incidence*, i.e., considering the prevalence of social media evidence, systems may prioritize specific items while disregarding others in their efforts to comprehend the psychological aspects of users on social media. Additionally, our interest extends to investigating whether the quantity of evidence related to BDI (Behavior, Depression, and Interaction) items fluctuates across different social media platforms and whether this fluctuation is linked to the mental health condition of individuals or is a result of the inherent characteristics of each social media source.

This study comprehensively examines the 21 BDI items and their impact on social media. We aim to explore the feasibility of deducing responses to this psychological test through automated methods. To achieve this objective, we initially establish a categorization system based on available evidence for each BDI item from social media (i.e., their incidence). This helps group the 21 items into different categories. Next, we examine the occurrence in conjunction with additional characteristics to comprehend the factors contributing to estimating a BDI item. Our main research questions (RQ), therefore, are:

- **RQ1**: For individuals in the positive group experiencing depression, to what extent can we access information related to each BDI item from their social media feeds?
- **RQ2**: To what degree do the levels of item occurrence (and its potential characteristics) correlate with the success achieved by automated techniques that deduce a response to the item using data from social media?
- **RQ3**: How extensive is the information present on the feed of control users regarding each BDI item when compared to positive users[2]?

To our understanding, this is the first study examining the incidence of BDI items on social media. This examination represents a significant stride in enhancing risk assessment tools and aiding decision-making. Our main findings reveal that certain items show a more sustained prevalence than others and, thus, have potentially stronger pieces of evidence. However, more information about these topics also brings more noise. This observation is corroborated by the sub-optimal performance

[2] Throughout this work, we utilize the word *positive* to refer to individuals who have been diagnosed with depression, while *control users* are those who are labeled as not suffering from this disorder.

of many eRisk participating systems, which tried to deduce the answers to the BDI by analyzing users' posts.

The primary aspects of our contributions can be outlined as follows:

1. A detailed study of the incidence of specific depression symptoms on social media. To do this, we used the BDI questionnaire as a reference clinical inventory.
2. An analysis of how standardized symptoms of depression affect the effectiveness of automatic screening tools. We take special attention to the existence of evidence about these topics and their correlation with other features.

The rest of the document is structured as follows. Section 2 provides an overview of the relevant literature, while Sect. 3 outlines the methodology used to address the research questions. Section 4, the data utilized in this study is discussed, and Sect. 5 showcases the experiments and analyses conducted. Finally, Sect. 7 presents conclusions and potential avenues for future research.

2 Related work

There is increasing interest in the scientific community to analyze the narrative impact of depression in social media or forums [34], and many studies have used self-reported inventories, like BDI, to select users from these platforms. For example, De Choudhury et al. [31] conducted initial research on the automated detection of depression by collecting data on depression prevalence through crowd-sourcing methods. They adopted the CES-D (Center for Epidemiologic Studies Depression Scale) [35] inventory to quantify the depression levels of the participants and requested them to grant access to their Twitter public feeds. Similarly, the CES-D self-assessment questionnaire was exploited by Reece et al. [36], who collected several Twitter users who were diagnosed with either depression or post-traumatic stress disorder (PTSD). The study conducted a state-space temporal analysis whose goal was to track the evolution of these disorders. Furthermore, Chu et al. [37] explored the use of different questionnaires, such as the Family Resilience Assessment Scale (FRAS), Caregiver Burden Inventory (CBI), Beck Depression Inventory-II (BDI), and Behavioral Measures Questionnaire, and combined them with Bayesian networks to assess the effects of family resilience on caregiver behavioral problems. This work enhanced the comprehension of the connections between depression, caregiver stress, and family resilience using these clinical instruments.

Skaik et al. [38] utilized the eRisk 2021 Task 3 dataset [24] to develop an automated system aimed at completing the Beck Depression Inventory (BDI) questionnaire for some Reddit users. Their approach involved building specific BDI question models and selecting top-performing models for each question. Subsequently, these models were combined into a single model named "BDI-Multi-Model". The model outperformed the existing state-of-the-art solutions for this task. Following this, the authors tried to transfer the model to a dataset of the Canadian population. To do this, this team compared the predictions generated by their model with the

statistics of the most recent mental health survey conducted by Statistics Canada. The results demonstrated a strong Pearson correlation of 0.90 between the questionnaire responses generated by their model and the official statistical data. Moreover, Kang et al. [39] introduced an advanced deep learning model to forecast the BDI score by leveraging electroencephalogram (EEG) data. This study demonstrated the feasibility of accurately anticipating BDI scores using EEG data. Schwartz et al. [40] utilized various predictors to assess the neuroticism personality trait. Their scales matched self-reported items designed to identify depression [41]. These authors investigated the correlation between the level of depression (severity) and the language patterns of a specific group of Facebook users. The degree of depression for each user was determined by calculating the average response to seven items related to depression.

Park et al. [42] investigated the possibility of effectively interpreting self-report inventories based on language usage on social media. To explore this, participants were asked to complete a depression self-assessment questionnaire and granted permission for a single extraction of their Twitter feed. The analyses conducted indicated a noteworthy distinction in the use of words conveying anger and negative emotions between the groups categorized as depressed and those classified as non-depressed.

A few studies have explored depression symptom discovery based on self-assessment questionnaires and DSM-V[3] [43] instruments. Gaur et al. [44] introduced an unsupervised method for aligning the content found in different mental health-related subreddits with corresponding DSM-5 categories. Their approach involved creating a specialized lexicon specific to the domain, incorporating n-grams linked to each mental health disorder within the DSM-5. This lexicon was developed by leveraging the DSM-5 manual and other carefully curated medical knowledge bases. The authors also expanded an ontology related to drug abuse, enriching it with mental health-related terminology and slang terms extracted from Reddit. Subsequently, these lexicons were used to assess the association between the content of the subreddits and the DSM-5 categories.

Similarly, Yazdavar et al. [45] created a lexicon for depression that consists of prevalent symptoms of depression sourced from the clinical assessment inventory PHQ-9 [46]. They leveraged the lexicon as background knowledge to compile a set of seed terms for each symptom. Using the seeds, they trained a semi-supervised topic model over the tweets of a set of users who self-declared as depression sufferers. The model was exploited to derive depression symptom distributions and word distributions that help to screen symptom trends over time. In a more recent study, Gu et al. [47] introduced a dataset gathered from Sina Weibo and addressed identifying depression as a binary classification challenge. By leveraging domain expertise in depression and the Dalian University of Technology Sentiment Lexicon (DUT-SL), they were able to build their own specialized depression lexicon with enhanced linguistic characteristics. Moving forward, in [48], the authors sought to predict individuals experiencing depression and gauged the severity of their depression by

[3] DSM-5 stands for *Diagnostic and Statistical Manual of Mental Disorders-5th Edition.*

analyzing data from Twitter. These authors employed a self-supervised approach to provide preliminary labels for the Twitter data. The approach was based on the extraction of user characteristics, including emotional, thematic, behavioral, user-specific, and depression-associated attributes. Using these attributes, they developed a long short-term memory (LSTM) network to make predictions regarding the intensity of depression. In a more recent work [49], the authors created models for detecting suicidal ideation by employing both traditional machine learning and advanced deep learning techniques, such as Transformers. These models were tested on three distinct datasets related to suicide detection.

Social media not only provides textual information but also rich metadata information. Related to this, Ghosh et al. [50] proposed that Twitter metadata can offer valuable indicators of depression, facilitating early assessment. This team created a comprehensive, multi-modal, multi-task system for detecting depression as the primary objective and recognizing emotions as a secondary task. This system leverages diverse emotional expressions within user descriptions to enhance the primary task's learning process. Additionally, a highly innovative approach was proposed by Sadasivuni et al. [51], who endeavored to analyze tweets concerning depression and anti-depression topics. They introduced a novel parameter that gauges the level of depression on a given day. When contrasted with historical data, this parameter becomes a valuable resource for social scientists interested in examining the effectiveness of psychotherapy and the prevalence of depression. The authors' method aimed to contribute to advancing mental health, psycho-social interventions, and pursuing sustainable development goals.

In recent work, Belcastro et al. [52] introduced an approach for effective and interpretable analysis of depression from social media posts. Their work offers a step forward by leveraging interpretable AI techniques to provide clear insights alongside depression predictions. Specifically, the potential integration of Large Language Models (LLMs) holds promise for enhancing the explainability of this approach.

Overall, considerable effort has been dedicated to identifying and predicting depression symptoms from online sources (see the recent survey [10]). However, research on the prevalence of individual sub-items of self-assessment inventories, like the BDI, and on the characteristics of text that facilitate effective estimation of depression has been noticeably scarce. In our research, we are specifically interested in studying and gaining new insights about the incidence of various symptoms linked to depression on social media. We are highly committed to understanding the feasibility of mining BDI symptoms from online sources.

3 Objectives and methodology

In this section, we describe the experiments and the methodology designed to study the presence of the 21 BDI items on social media posts. The objective of these experiments is to address the research questions presented in Sect. 1. We outline

the insights we can gain from each experiment, emphasizing the connection to each research question.[4]

3.1 Background

This research focuses on analyzing depression, a mental health condition characterized by persistent sadness and diminished interest in previously enjoyed activities. It can also manifest in changes in sleep patterns, appetite, and energy levels. Although occasional sadness is a normal part of life, depression is different because it is prolonged and significantly disrupts daily functioning.[5] We aim to assess various aspects of depression using the well-established Beck Depression Inventory-II (BDI). This standardized psychological questionnaire, consisting of 21 multiple-choice questions, delves into different facets of the subject's emotional state. The BDI is self-administered, meaning individuals answer the questions themselves, selecting options with scores ranging from 0 to 3. A scoring guide at the questionnaire's end translates the responses into an overall score. As an example, the "sadness" section presents the following answer choices:

- 0—I do not feel sad.
- 1—I feel sad.
- 2—I am sad all the time, and I can't snap out of it.
- 3—I am so sad and unhappy that I can't stand it.

3.2 BDI item incidence

To address **RQ1** and **RQ3**, we assess the impact of each BDI item on social media by quantifying its *incidence*. Essentially, we gauge the level of information available about each item in a set of social media posts. We assume the available evidence differs for each item, enabling us to rank and compare the BDI items using estimates of their respective incidences.

Figure 1 depicts an overview of the methodology. We start by pre-processing a collection of social media posts and transforming them into index terms (e.g., words). To do this, we use a Python library called *ekphrasis* [53]. This library is tailored for text from social media sites. It performs tokenization, word normalization and segmentation (e.g., for splitting hashtags), and spell correction. In a subsequent step, the various terms alongside additional information (e.g., term frequency) are organized into data structures that enable fast searching. Next, a query is formed from each BDI item in the questionnaire. Using a ranking function, we retrieve relevant posts from the inverted index for each query and compute an incidence level based on the relevance score of the top-k documents. Finally, we categorize BDI items according to their estimated incidence.

[4] Code available at: http://bit.ly/3xHEolB.

[5] https://www.euro.who.int/en/health-topics/noncommunicable-diseases/mental-health.

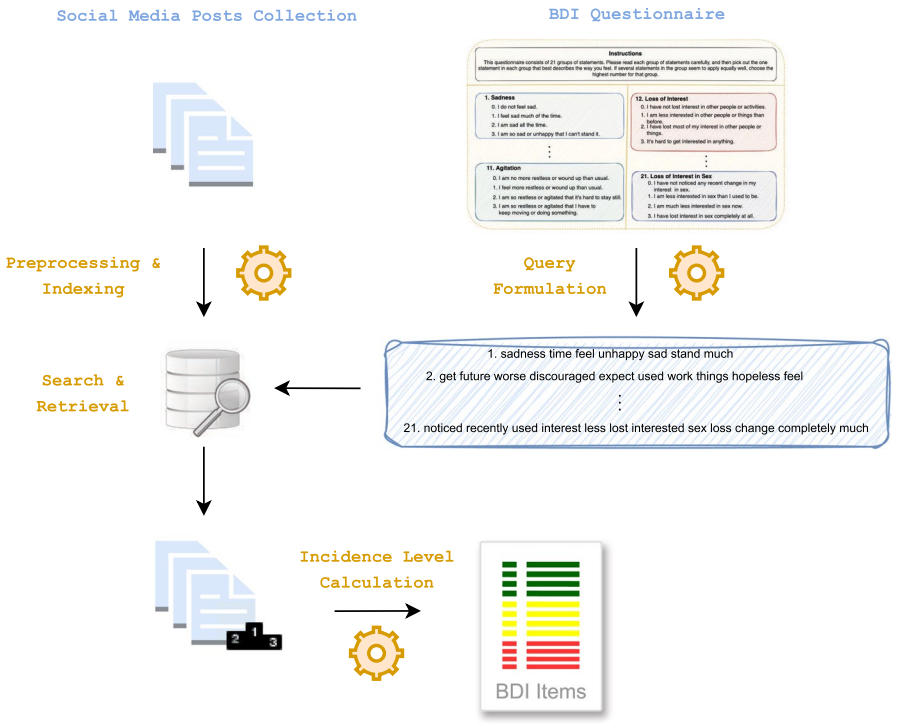

Fig. 1 General diagram of the proposed methodology. The overall goal is to estimate incidence levels on social media posts using queries generated from the BDI questionnaire

3.2.1 Incidence level

Formally, let s_{ijk} represent an *incidence* score, calculated as the mean relevance score of documents (referred to as social media posts in this article) retrieved for BDI item q_i using the ranking function f_j from collection c_k. Using these incidence scores, we can arrange the BDI items to form a ranked list comprising 21 elements. This ranked list will be called r_{jk}, and the position of the $i-th$ BDI item on this list will be denoted as p_{ijk}. Next, we split the ranked list r_{jk} into four groups. Each group signifies the amount of evidence discovered, as measured by s_{ijk} compared to other items in the questionnaire. To elaborate, we establish $I(q_i)$ as a function that designates an *incidence level* (or category) to each BDI item based on its placement p_{ijk} within the ranking r_{jk}:

$$I(q_i) = \begin{cases} High\ (+), & \text{if } p_{ijk} \geq 1 \ \wedge \ p_{ijk} \leq 5 \\ Middle\text{-}High\ (\wedge), & \text{if } p_{ijk} \geq 6 \ \wedge \ p_{ijk} \leq 10 \\ Middle\text{-}Low\ (\vee), & \text{if } p_{ijk} \geq 11 \ \wedge \ p_{ijk} \leq 15 \\ Low\ (-), & \text{if } p_{ijk} \geq 16 \ \wedge \ p_{ijk} \leq 21 \end{cases}.$$

The rationale behind categorizing the items into four groups is to ensure a relatively equal distribution of items across each level of incidence, except for the last

category, which contains one additional item. This classification aids in the analysis of BDI items by examining their proportional presence on social media platforms. Items with a high incidence are associated with more online evidence compared to other items in the psychological questionnaire. Ideally, automated screening tools would benefit from the increased availability of relevant information. Understanding incidence levels could improve the construction of depression screening systems. For instance, such systems could prioritize items with higher incidence, allowing for more extensive information extraction while disregarding items with limited evidence. This selective approach may contribute to more reliable estimates in depression screening. In addition, delve into the language and conduct of individuals impacted by mental disorders can uncover novel predictive indicators that have not been previously considered in the medical literature. This would be promising for designing new psychometric surveys and questionnaires (e.g., considering recent investigations into the behavioral characteristics associated with mental health disorders [2]).

It is crucial to emphasize that the approach described here represents one of several potential avenues to classify and assess the occurrences of BDI items on social media. For example, another alternative could be to compute incidence scores from the average relevance scores produced by the retrieval system and, next, define a set of thresholds and incidence levels accordingly. The limitations related to the methodological choices adopted in this work are discussed in Sect. 6.

3.2.2 Search methods

To perform information retrieval experiments, we utilize Okapi-BM25 (BM25) [54] and Query Likelihood with Dirichlet Smoothing (QLD) [55], which are two well-established ranking methods in the field of Information Retrieval [56]. BM25 is a lexical matching score that estimates the degree of relevance of a document with respect to an input query. BM25 implicitly represents documents and queries as "bags of words" and produces an overall retrieval score from accumulating individual document-query matching weights. The matching words' weights incorporate a saturated term frequency (TF) component, which grows with the number of occurrences of the query word in the document, an inverse document frequency (IDF) component, which gives more importance to rare words, and a document length normalization component. Although classic, BM25 is a well-known reference in search technologies. It is still commonly used in state-of-the-art solutions (e.g., as an initial retrieval stage that feeds candidate documents to neural retrieval systems [57]). The Query Likelihood model is another classic retrieval approach. It is based on statistical language models and estimates the likelihood of generating the query from a statistical language model built from the document's text. Under this model, every document in a collection is regarded as a language model, while a user's query is treated as a sequence of words. In essence, the model quantifies how probable it is for a document to generate the words in the user's query. QLD is a particular instance of query likelihood models that boils down to a retrieval matching function with standard IR components (TF-like, IDF-like, and length normalization weights).

Both models, QLD and BM25, are competitive lexical matching search methods. More details about them can be found elsewhere [56].

3.2.3 Document collections

This study uses document collections from two popular social media platforms, e.g. Reddit and Twitter. This approach allows the examination of two distinct social media platforms with unique features and limitations. Our focus is primarily on assessing the consistency of trends and patterns observed across both sources, irrespective of their specific characteristics. The occurrence score is calculated using document relevance scores for each BDI item through post-level analysis. We employ two levels of analysis, depending on the number of ranked documents analyzed (k): *global-incidence* ($k = 1000$) and *top-incidence* ($k = 10$). The scores in these ranked lists are averaged to obtain the item's incidence estimate (s_{ijk}). Using this approach, we can examine incidence at various levels.

Guan et al. [58] examined the impact of ranked list positions on navigational and informational search tasks, discovering that users tend to favor top-ranked results, irrespective of their relevance. In particular, Google's initial results page captures the attention of 95% of Internet users.[6] In the context of most contemporary commercial web search engines, result pages typically consist of 10 elements. The highest incidence score aims to capture the prominence of some social media submissions that prove highly relevant to the specified BDI item, regardless of how closely the BDI item aligns with the overall collection of documents. We chose $k = 10$ because this setting represents the standard first page of the result in current retrieval engines. The global incidence approach provides a broader perspective on the presence of the BDI item within a document collection. To ensure comprehensive coverage of relevant documents, we employed $k = 1000$. A high top incidence score for a specific BDI item indicates that a few users extensively discussed the topics and used words associated with the BDI item. However, a high global incidence score suggests a more sustained prevalence of the BDI item across the entire collection of social media submissions, which means that evidence related to the item is spread over a larger set of documents. Our interest lies in examining the relative differences in top and global incidence for all BDI items.

3.2.4 Query formulation

For the retrieval experiments, we built a textual query for each BDI item. In this process, we examined each questionnaire item and formulated a query by combining the item's title with the content of all possible responses. As Croft et al. [56] explained, removing stopwords generally improves retrieval effectiveness; therefore, common words were eliminated during this step utilizing the NLTK's stoplist.[7] The decision to extract query words from both the title and responses

[6] Refer to http://bit.ly/38OKy4z.

[7] Available at https://gist.github.com/sebleier/554280.

Table 1 Queries generated for each BDI item (after concatenating the text of the title and the available responses and removing stopwords)

#	Queries
1	Sad much feel sadness time unhappy stand
2	Used pessimism get work worse discouraged things hopeless future expect feel
3	See person past failures feel total back look lot failed failure
4	Enjoy much pleasure get little things ever used loss
5	Guilty quite done particularly time feelings things many feel
6	Punished feelings expect punishment feel may
7	Self-dislike confidence feel disappointed ever lost
8	Blame bad everything self-criticalness criticize faults critical happens used usual
9	Suicidal kill thoughts wishes carry myself chance killing
10	Little every cry thing feel crying anymore used
11	Moving keep hard still feel usual agitation agitated stay something wound restless
12	Hard less things activities loss interested interest get anything lost
13	Difficulty decisions find much greater ever making indecisiveness well usual trouble difficult used make
14	Worthlessness used feel worthwhile consider utterly compared worthless useful
15	Used energy loss much anything ever enough less
16	Usual wake changes experienced less lot pattern hours sleep 1–2 day sleeping back change early somewhat get
17	Irritability much time usual irritable
18	Usual timeless appetite much greater crave changes experienced somewhat food change
19	Concentration mind find keep anything hard well ever usual difficulty long concentrate
20	Tired things usual lot easily used fatigue fatigued tiredness get
21	Recently noticed lost less completely used sex change much interest interested loss

was intentional, as the title captures the overarching theme of the item (e.g., sadness). In contrast, the responses offer varied expressions of how individuals may feel about this symptom of depression. Preliminary experiments using short queries constructed only from an item's title revealed that people rarely employ general expressions (such as sadness, agitation, or indecisiveness) to convey their emotions. Consequently, we constructed more extensive queries by incorporating all available words (title + responses). The resulting set comprises 21 queries, as detailed in Table 1. While applying lexical word-matching search strategies may potentially overlook indirect indicators about BDI items within user-generated posts, such as function words, this straightforward yet highly precise approach serves as a baseline for future experiments. In addition, these conventional strategies have demonstrated their efficacy in traditional search scenarios. We are not interested in optimizing retrieval performance, but in comparatively analyzing the presence of BDI-related content on social media. In the conclusion, we discuss more sophisticated query construction methods and alternative search strategies based on, for example, semantic similarity.

3.3 Depression severity estimation

In response to **RQ2**, it is essential to employ specific metrics to evaluate the efficiency of systems that derive BDI responses from social media data. To achieve this objective, we utilize two metrics endorsed by Losada et al. [22, 23]. These metrics assess the alignment between the BDI questionnaire responses provided by an actual social media user and those generated by a system based on the user's posts. The comparison involves examining how closely the system's questionnaire aligns with that of the real user within a given set of social media users:

- Average Hit Rate (AHR) is the averaged Hit Rate (HR) across all users. HR serves as a strict metric, calculating the proportion of instances where the automatically generated questionnaire matches the responses of the actual questionnaire.
- Average Closeness Rate (ACR) is the mean Closeness Rate (CR) computed over all available users. CR considers the ordinal nature of the responses to the BDI questionnaire and provides a performance estimate based on the distance between the real and estimated responses.

We analyze the correlation between the effectiveness of automated systems that respond to the BDI questionnaire and specific characteristics of the BDI, such as the presence of the BDI item on social media. To achieve this, we examined the submissions made by multiple research teams to eRisk 2019 and eRisk 2020. These were 20 and 18 variants (also called runs), respectively. In essence, participants of this experimental shared-data task were given the history of each user's posts. They had to design algorithms that extract useful depression-related evidence and subsequently predict how the user would respond to each BDI item. Various systems have been proposed to face this challenge. In this analysis, we graphically represent each BDI item, including its AHR and ACR (averaged across all systems involved), the BDI item's incidence estimates, and various other factors. Additionally, we calculate Pearson's correlation coefficient to assess the correlation between the analyzed variables.

4 Data

In this study, we worked with the collections published in eRisk 2018 [21] (Reddit) and CLPsych 2015 [27] (Twitter). These initiatives provided a common evaluation framework to foster research on the automated recognition of mental disorders through online social media platforms. It should be noted that this research represents an initial attempt in the direction of comprehending the incidence of depression symptoms on social media. In the future, we plan to extend this analysis to additional online sources. This would help to better understand how many mental health screening tools can be generalized across multiple social media platforms.

Each collection consists of entries submitted as posts by individuals on social media platforms, specifically Reddit and Twitter. Each dataset contains two distinct

user groups: (a) positive group, composed of users potentially with depression; and (b) control individuals.

In both experimental setups, participants in the positive group were gathered by searching each platform for self-disclosures of diagnoses (such as, "I have been diagnosed with depression") and ensuring through manual verification that the retrieved posts indeed included authentic statements of diagnosis [59]. Conversely, individuals in the control group were chosen randomly from the vast user base of each platform.

It is important to note that the organizers of eRisk aimed to obtain a comprehensive understanding of users' language. As a result, the eRisk collections encompass a variety of discussions and issues that span a wide spectrum of topics. All user's submissions, regardless of the specific subreddit (Reddit's topic-based communities) where they were posted, were incorporated into the collection. Moreover, the organizers also included in the control group users who do not suffer from any disorder but participate in Reddit forums focused on depression or mental well-being. This type of user includes, for example, clinical practitioners offering assistance to individuals at risk or subjects who have a family member experiencing depression. These users with a supportive role make the eRisk goal (i.e., distinguishing positive and control group users) becomes more challenging. This is because the topics of interest of these control group users are highly related to those of the positive group. However, we are not interested in a two-class categorization problem but in the incidence of specific BDI items over these collections. Thus, we opted to build a less noisy control group of users. In particular, to tackle **RQ3** and mitigate potential biases, we opted to replace the 2018 control group with the 2019 control group of eRisk [22].[8] In this setting, users who give support or are active in the mental health subreddits are excluded from the control group.

Due to the API constraints set by each social media platform, the dataset creators managed to gather the most recent 2000 submissions per user for eRisk and 3000 for CLPsych. A concise overview of these data collections is presented in Table 2.

5 Experiments and analysis

To conduct the various retrieval experiments, we used Anserini[9] [60], an open-source information retrieval toolkit. We indexed each collection and group of users separately, obtaining four different indexes. Retrieval was performed by issuing to Anserini the queries constructed from each BDI item. Query production followed the strategy outlined in Sect. 3. We used the BM25 and QLD ranking functions as implemented in Anserini (with the default parameter settings). It is worth noting that we consider individual posts or tweets as the retrieval units for this study. Each post or tweet in a particular collection is considered a single document. Hence, the more

[8] The 2019 edition of eRisk was focused on the early detection of signs of self-harm.

[9] See https://github.com/castorini/anserini.

Table 2 Summary of eRisk's (Reddit) and CLPysch's (Twitter) collections

	Reddit		Twitter	
	Positive	Control	Positive	Control
# of Users	214	299	522	872
# of Documents	89,999	161,886	1,131,997	1,978,121
Avg. # of Documents/User	541.42	658.2	2,373.73	2,286.51
Avg. # Words/Document	45.0	28.9	13.9	13.8
# of Unique Words	41,986	70,229	150,508	238,712

individual documents that are on-topic, the higher the incidence of the BDI item across social media publications.

5.1 BDI item incidence

We begin by examining the occurrence of BDI items for each user group separately, distinguishing between the positive and control groups. Subsequently, we compare the incidence trends between these two groups.

5.1.1 Positive groups

Table 3 compares the top and global incidence for positive users in Reddit and Twitter. In general, we note that there are discernible variations in the occurrence of different BDI items on social media platforms. At the global scale, specific elements like 13 (*indecisiveness*) and 16 (*sleeping patterns*) consistently demonstrate similar trends across various ranking methods and datasets. This indicates a higher frequency of discussions and greater evidence that can be gathered and examined regarding these subjects. In contrast, elements 1 (*sadness*), 6 (*punishment feelings*), and 17 (*irritability*) display contrasting patterns, indicating a scarcity of available evidence for these items. On Twitter, aspects like 7 (*self-dislike*) are more noticeable than Reddit's. Interestingly, item 7 and item 8 (*self-criticalness*) exhibit a higher occurrence at the top level on Reddit, exceeding their overall prevalence. This indicates that on Reddit, these two themes were not widespread across all users, but a select few posted highly relevant content. Similarly, item 11 (*agitation*) has a moderate global presence on Twitter, but its prevalence at the top level on this platform is significantly higher. It is important to include a word of caution here. If an item has a low incidence level, it does not necessarily mean that there is no social media evidence related to it. Rather, it indicates that existing evidence is not readily accessible through standard search methods. This retrieval limitation poses a barrier for systems that seek to learn about the symptoms associated with the BDI item. In contrast, obtaining documentary evidence is not complicated for items with higher incidence.

Table 3 BDI item's incidence level: High (+), Middle-High (^), Middle-Low (v), and Low (−) incidence categories. Comparison of positive groups at global and top scales. *R* stands for Reddit and *T* for Twitter

#	BDI Item	Global ($k = 1000$)				Top ($k = 10$)			
		$BM25_R$	QLD_R	$BM25_T$	QLD_T	$BM25_R$	QLD_R	$BM25_T$	QLD_T
1	Sadness	−	−	−	−	−	−	−	−
2	Pessimism	+	v	+	−	+	^	v	^
3	Past Failure	^	v	^	^	^	v	^	^
4	Loss of Pleasure	^	−	^	v	v	−	v	−
5	Guilty Feelings	^	v	^	v	^	v	^	−
6	Punishment Feelings	−	−	−	−	−	−	−	v
7	Self-Dislike	v	^	+	+	+	+	+	+
8	Self-Criticalness	−	^	−	^	v	+	v	+
9	Suicidal Thoughts	−	^	−	^	−	^	v	v
10	Crying	+	+	+	−	+	+	+	−
11	Agitation	v	−	^	v	v	v	+	+
12	Loss of Interest	^	^	v	^	v	−	^	^
13	Indecisiveness	+	+	+	+	^	^	^	v
14	Worthlessness	−	^	−	+	−	v	−	−
15	Loss of Energy	v	−	v	−	−	−	−	−
16	Sleeping Patterns	+	+	+	+	+	+	+	^
17	Irritability	−	−	−	−	−	−	−	+
18	Changes Appetite	v	+	v	+	v	^	v	v
19	Concentration	+	v	^	v	+	v	+	^
20	Tiredness/Fatigue	v	v	v	v	^	+	−	+
21	Loss Interest Sex	^	+	v	^	^	^	^	v

5.1.2 Control groups

Table 4 presents a comparison of incidence levels on a global and top scale between two control user groups (Reddit and Twitter). Similarly to the positive groups, items 13 (*indecisiveness*) and 16 (*sleeping patterns*) demonstrate a high global incidence in various retrieval models and collections. Conversely, items 15 (*loss of energy*) and 17 (*irritability*) consistently exhibit low incidence levels, both globally and locally. Items 2 (*pessimism*), 19, and 21 (*loss of interest in sex*) show an overall lower incidence compared to the positive user groups. Notably, item 10 (*crying*), which might be expected to be highly discussed by positive users, displays similar patterns in both control groups, and all users consistently express concerns about this item.

5.1.3 Positive versus control groups

Figure 2 displays a boxplot illustrating the incidence scores of BDI items across the two collections and user groups. The primary objective of this analysis is to examine how incidence scores vary across different collections and groups. To achieve this, we categorized BDI items into three groups based on their consistent high, moderate, or low incidence. Following Fuhr's suggestion [61], we calculated the effect size using Cohen's D [62] and assessed statistical significance using the Welch two-sample t-test with a *p*-value of less than 0.001. Table 5 presents the effect sizes for the three groups of BDI items.

Initially, when examining the positive groups on Twitter and Reddit, it becomes apparent that, on average, the documents obtained from Twitter exhibit a higher

Table 4 BDI item's incidence level: High (+), Middle-High (^), Middle-Low (v), and Low (−) incidence categories. Comparison of control groups at global and top scales. *R* stands for Reddit and *T* for Twitter

		Global ($k = 1000$)				**Top** ($k = 10$)			
#	BDI-Item	$BM25_R$	QLD_R	$BM25_T$	QLD_T	$BM25_R$	QLD_R	$BM25_T$	QLD_T
1	Sadness	v	−	−	−	−	v	−	−
2	Pessimism	^	−	^	−	−	^	^	+
3	Past Failure	+	^	v	^	^	^	^	+
4	Loss of Pleasure	^	v	^	v	v	−	^	−
5	Guilty Feelings	^	v	^	v	v	v	v	−
6	Punishment Feelings	−	−	−	−	−	v	−	v
7	Self-Dislike	−	^	+	+	+	+	+	+
8	Self-Criticalness	−	^	v	+	^	^	v	+
9	Suicidal Thoughts	−	^	−	v	−	v	−	^
10	Crying	+	+	+	^	+	+	+	v
11	Agitation	v	v	^	^	v	^	+	^
12	Loss of Interest	^	^	v	v	v	−	^	−
13	Indecisiveness	+	+	+	^	^	−	+	^
14	Worthlessness	−	v	−	+	−	v	−	v
15	Loss of Energy	v	−	−	−	v	−	−	−
16	Sleeping Patterns	+	+	+	+	+	+	+	^
17	Irritability	−	−	−	−	−	−	−	+
18	Changes Appetite	v	+	v	+	^	+	v	^
19	Concentration	+	v	+	^	+	^	^	v
20	Tiredness/Fatigue	v	v	^	v	^	+	v	v
21	Loss Interest Sex	^	+	v	v	+	−	v	−

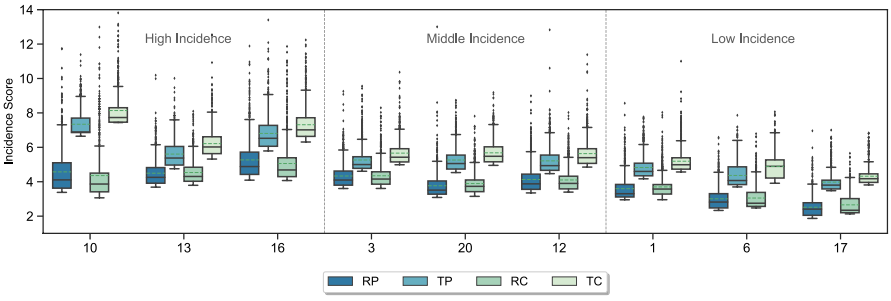

Fig. 2 The incidence scores of BDI items for both the positive and control groups are depicted in the distribution. In this context, RP refers to Reddit Positive, RC refers to Reddit Control, TP refers to Twitter Positive, and TC refers to Twitter Control

relevance score than those obtained from Reddit. Specifically, the identified disparities, as shown in Table 5 (in columns designated RP vs. TP and TC vs. RC), can be characterized as having a "very large" effect size ($\Delta > 1.2$) [62]. This implies that there is substantial evidence, in the form of tweets compared to Reddit posts, about the topics of the BDI items. A possible reason for this could be related to the distinctive features and limitations of the social media platforms examined, such as spatial restrictions. Twitter users are restricted to a limited character count, whereas Reddit users tend to compose comparatively longer texts on average (see Table 2). Tweets represent concise and focused messages, and if they align with a specific BDI item, their relevance score is not significantly penalized when considering length normalization. Another reason could be the difference in the scale of the examined datasets.

Table 5 Effect sizes and p-values calculated based on the incidence scores depicted in Fig. 3, the abbreviations RP, RC, TP, and TC correspond to Reddit Positive, Reddit Control, Twitter Positive, and Twitter Control, respectively

#	RC versus RP		TC versus TP		RP versus TP		RC versus TC	
	Δ	ρ-value	Δ	ρ-value	Δ	ρ-value	Δ	ρ-value
10	0.1511	0.0007	0.8984	≈ 0	2.5979	≈ 0	2.9270	≈ 0
13	0.0507	0.2568	0.7558	≈ 0	1.4040	≈ 0	2.1710	≈ 0
16	0.1767	≈ 0	0.5193	≈ 0	1.3933	≈ 0	2.1344	≈ 0
3	0.0200	0.6543	0.5474	≈ 0	1.2608	≈ 0	1.8440	≈ 0
20	0.1708	0.0001	0.6113	≈ 0	2.0717	≈ 0	2.5042	≈ 0
12	0.0340	0.4475	0.5279	≈ 0	1.3584	≈ 0	2.0211	≈ 0
1	0.1478	0.0009	0.5347	≈ 0	1.7741	≈ 0	2.1913	≈ 0
6	0.0673	0.1323	0.7303	≈ 0	1.9533	≈ 0	2.6199	≈ 0
17	0.2294	≈ 0	0.6469	≈ 0	2.5242	≈ 0	2.9391	≈ 0

The Twitter dataset is nearly twelve times larger than the Reddit dataset, enhancing the likelihood of obtaining pertinent content. To shed light on this issue, we plan to conduct future experiments against Twitter and Reddit APIs to further validate these findings.

Secondly, we observe a notable pattern in the incidence scores of both the positive and control groups. Upon examining the effect sizes between the eRisk positive and control groups (RP vs. RC, as shown in Table 5), we observe minimal effect sizes, and none of the differences are statistically significant in any instances. In contrast, we find larger effect sizes when comparing the Twitter groups (TP vs. TC, as indicated in Table 5). Notably, the Twitter control group exhibits a marginally higher incidence score on average than the positive group. However, this pattern is not observed on Reddit. One possible explanation for this finding is that the topics addressed by the BDI items are widely discussed regardless of the users' mental health status. Nevertheless, as previously demonstrated, most distinctions between positive and control groups are measurable in how individuals articulate language and display emotional cues rather than the extent to which diverse themes are discussed.

5.2 Depression severity estimation

Figure 3 illustrates the incidence score of BDI items and the mean ACR and AHR achieved by the systems participating in eRisk 2020.[10] The negative correlation

[10] Similar findings were observed in eRisk 2019; however, to prevent unnecessary overload of figures, we present ACR and AHR plots only for eRisk 2020.

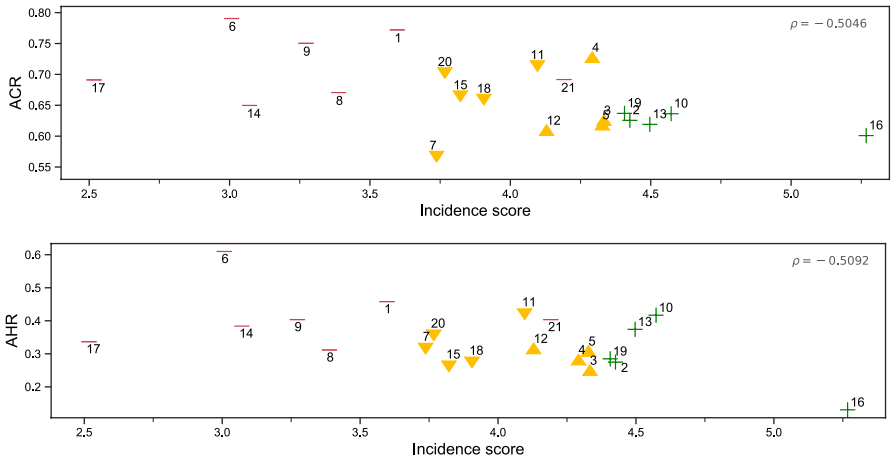

Fig. 3 The diagram illustrates the relationship between incidence, calculated using BM25 on Reddit, and average effectiveness, measured by ACR and AHR using the eRisk 2020 systems. The report includes the Pearson's correlation coefficient, with significance determined through a Bonferroni correction at a p-value of $< 0.025, (0.05/2)$. The incidences are categorized as High (+), Middle-High (▲), Middle-Low (▼), and Low (−)

is evident in Pearson's correlation coefficients, which are −0.5046 for ACR and −0.5092 for AHR. After applying the Bonferroni correction, both coefficients are statistically significant at a p-value of $< 0.025, (0.05/2)$. This implies that as the amount of evidence increases, the efficiency of the systems diminishes. For example, elements such as "punishment feelings" and "sadness" (refer to items 6 and 1 in Table 3) exhibit elevated ACR but low incidence scores. In contrast, items with high incidence, such as "sleeping patterns" (item 16), produce a low ACR. The more relevant documents for a specific BDI item, the more noise is introduced. Our analysis indicates that current systems struggle to detect irrelevant and noisy information.

We analyzed the length of queries and found that items with shorter representations tended to be more effective. This observation was supported by significant Pearson's correlation coefficients of −0.5228 for AHR and −0.4370 for ACR. Longer queries derived from BDI items introduced more off-topic terms, leading to a phenomenon known as query drift. Recall that the query words were extracted from the item's title and responses. These textual elements serve as the core source of information for systems to represent the BDI item. Most participants generated queries or representations from the questionnaire without undergoing additional post-processing. Our correlation analysis, examining the relationship between query length and effectiveness, indicates that BDI items with a substantial length may potentially contain too much noise.

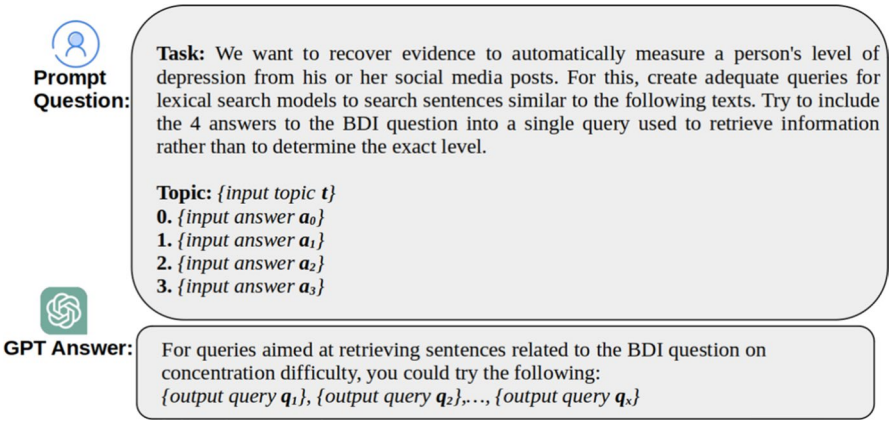

Fig. 4 Example of the prompt used to generate new queries. The text indicates the task, the corresponding topic, and the possible answers for each topic

We also examined the average inverse document frequency (IDF) for each query.[11] Essentially, IDF gauges how uncommon a word is within a corpus. The objective of the IDF analysis was to determine whether the distinguishing significance of the terms derived from the BDI items influenced the effectiveness of the systems. However, we did not observe a correlation between the average IDF and system performance.

5.3 Query formulation using large language models

In recent years, Large Language Models (LLMs) have received significant attention, particularly since ChatGPT[12] emerged. This is owing to their remarkable performance across various natural language tasks. The prowess of LLMs in comprehending and generating language stems from their capacity to self-train from vast amounts of textual data and their sophisticated neural network architectures, which can encode human knowledge in parametric form.

One potential limitation of our study lies in its query formulation strategy. Relying solely on words extracted from the title and responses of the BDI item does not necessarily capture the various ways in which depressive symptoms are expressed on social media. In this regard, we explore the capacity of LLMs to generate text, and more specifically, we employ them to build new queries for each BDI item. These LLM-based queries are run against the search models and we recalculate the incidence values of the BDI items accordingly. The main goal of this exploratory study is to probe whether the relative order of the BDI items, estimated from their incidence scores, is affected by the query formulation strategy.

[11] We utilized the MS MARCO Passage Retrieval collection to gather the term statistics necessary for IDF computation. Accessible at https://microsoft.github.io/msmarco/.

[12] https://openai.com/research/chatgpt.

Table 6 Exploratory study based on queries generated by an LLM. The comparison involves positive and control groups at a global scale for a subset of BDI items. *R* stands for Reddit and *T* for Twitter

		Positive Group				Control Group			
#	BDI-Item	$BM25_R$	QLD_R	$BM25_T$	QLD_T	$BM25_R$	QLD_R	$BM25_T$	QLD_T
13	Indecisiveness	⌄	−	⌄	⌄	⌄	⌃	⌄	⌄
16	Sleeping Patterns	−	⌃	−	⌄	−	⌄	−	⌄
12	Loss of Interest	⌄	−	⌄	⌄	⌃	−	⌄	−
20	Tiredness/Fatigue	−	⌄	−	⌃	−	−	−	⌃
6	Punishment Feelings	+	⌄	+	+	+	⌄	+	+
17	Irritability	+	⌃	+	⌃	+	⌄	+	⌄

To achieve this, we have devised an approach that instructs ChatGPT[13] to generate new queries for these experiments. We fed ChatGPT with a description of the task, the BDI item's topic, and possible responses. Figure 4 shows the prompt template and the corresponding outputs. Using this method, we aim to enrich the heterogeneity of BDI-related evidence searches, thereby extracting contents in more diverse contexts.

Following the procedure outlined in Sect. 3.2, stopwords were removed from the queries generated by the LLM, and the resulting words were used to create a single query per BDI item. In general, we observed that the queries produced were longer than those in Table 1. Moreover, we noticed that in certain cases, ChatGPT tended to drift from the instructions, and, for example, the output query only covered a specific or restricted part of the BDI item. There were even cases where the queries were somehow "biased", containing cue words, such as depression, explicitly linked to the disorder. Additional prompt engineering techniques could improve this [63]. Finally, as described in Sect. 5, the queries were issued to Anserini, and both BM25 and QLD were used to retrieve evidence for each BDI item.

In the original retrieval experiments reported in Tables 3 and 4, several items exhibited a fluctuating incidence over different search methods and social media platforms and therefore it was not possible to identify a trend. While certain items, such as *irritability* or *indecisiveness*, behaved consistently across collections and ranking functions. With this in mind, we selected a subset of BDI items for the LLM-based experiments and focused on any changes in their global incidence level. In particular, we picked two representatives from each incidence category based on the results obtained in Sect. 5. More specifically, we chose *indecisiveness*, *sleeping patterns*, *loss of interest*, *tiredness/fatigue*, *punishment feelings*, and *irritability*.

Table 6 presents the results of this exploratory incidence experiment using LLM-generated queries. We note that, for the positive groups, more evidence was retrieved for items 6 (*punishment feelings*) and 17 (*irritability*) than in the previous retrieval experiment (their incidence level has considerably increased). Conversely, elements 13 (*indecisiveness*) and 16 (*sleeping patterns*) show a lower incidence overall. One reason could be that the queries generated by ChatGPT for items 6 and 17 have broader coverage than those in Table 1, as they include several lexical variations

[13] We used the GPT version 3.5-turbo-0125.

about the same topic. Thus, more evidence was retrieved for such items. The queries for items 13 and 16 show a higher lexical overlap for both query formulation strategies. However, the incidence level of items 12 (*loss of interest*) and 20 (*tiredness/fatigue*) have not considerably changed. Similar conclusions can be drawn for the control groups.

These initial experiments with LLM-based queries could be further complemented with additional tests with the entire set of BDI items and alternative search models. Furthermore, we noted certain variations in the relative order of some of the selected items. The next natural step is to compare multiple query formulation methods to better understand their retrieval capabilities. For example, it is interesting to answer the question: does one method retrieve more evidence than the other? As queries become more comprehensive and general, is *noisier* evidence retrieved? In the near future, we are interested in analyzing these issues with the aid of health practitioners. For instance, they can be presented with documents retrieved following different query formulation strategies and asked to assess whether they see enough documentary evidence to estimate an answer for a particular BDI item.

6 Discussion

6.1 Limitations

This study has been limited to the analysis of user behavior on two social media platforms and has focused solely on depression. To analyze the transferability of our findings and enhance the utility of automated mental health screening tools, we want to explore in the near future a broader range of online sources, including forums and microblogs. Analyzing user behavior across diverse platforms would provide a more comprehensive understanding of how online behavior manifests itself under different environments. In addition, we are committed to expanding the scope of this research to include other mental health disorders, such as anxiety (using the GAD-7) and loneliness (using the UCLA Loneliness Scale). This would allow for the development of more versatile screening tools with the potential to identify a wider range of mental health concerns.

Another important limitation is that the population under study is restricted to self-disclosing users. This potentially excludes people who experience mental health problems but do not openly discuss their condition. Also, we cannot discard the presence of a small portion of false positives in the positive set of users. Although self-disclosure expressions provide valuable information, they should be interpreted with caution and complemented with additional contextual analysis to ensure an accurate assessment of individuals at risk. In our future work, we want to explore new ways to increase the population of individuals in the positive group.

In our experiments, the search for BDI-related evidence has been based on "classical" retrieval methods. While a purely keyword-based approach might have limitations, it represents a good starting point for retrieving evidence and for relatively comparing the incidence of different BDI items. This initial approach can be combined with other techniques (e.g., semantic matching models) to better understand

how depression manifests on social media. A natural continuation of this research involves exploring advanced search methods, such as those based on dense retrieval. Given the nature and mechanics of such methods, some steps followed in this work, such as stopword removal, are likely to be revisited.

It is also important to note the methodological limitations of the estimates of the incidence of depression. As mentioned above, this research offers a comparative analysis of the online evidence available for the BDI items. We selected an organization that allowed us to categorize BDI items into various incidence groups. This has facilitated the analysis, but the categorization of the BDI items could have been done differently. For instance, we could have directly employed the incidence scores (i.e., average relevance score), setting certain thresholds, and obtaining incidence levels accordingly.

We observed a limitation in our method related to the difference between two items belonging to adjacent categories. In particular, the difference between the retrieval scores of items from adjacent categories might be small (see Fig. 2). Thus, an item with a middle-high incidence could have also been considered to have a high incidence (or the other way around). However, given its position in the ranking, it ended up in a specific group. Yet, our method paints an overall picture of the relative trends among BDI items and how they reflect on social media.

Another limitation is related to granularity. We opted to conduct this investigation, focusing on individual posts or tweets as units for retrieval. Alternatively, an exploration of incidence at the user level could be pursued by consolidating all posts from a user into a unified document. Such a user-level approach might offer additional information to improve the understanding of observed incidence levels. For example, if an item exhibits a higher incidence, is it due to widespread discussion among many users, or is its incidence driven by a small number of highly active users repeatedly engaging with the topic? Although such an analysis could be informative, its granularity would be constrained by the number of users available in each dataset. Our focus in this study was to enhance the comprehension of extracting post-level evidential information from social media platforms and incorporating it into automated screening tools. This has the potential to help healthcare professionals perform a more thorough assessment of individuals, serving as a complementary approach to conventional diagnostic techniques. We believe that our framework represents an initial step towards addressing this issue by offering valuable insights into the constraints associated with online sources. We are optimistic that this research will pave the way for further exploration and prompt broader research inquiries.

6.2 Ethical concerns

Exploring studies with human subjects presents a delicate subject concerning the ethical handling of data and the privacy of individuals [64]. The intricate nature of mental health research requires a thorough examination of the potential advantages and drawbacks of the study.

This study aims to improve our understanding of detailed indicators of depression, offering potential advantages in understanding the nuances of this mental

health condition and the extent to which automated early-risk methods could successfully capture them. This analysis is valuable for discovering the shortcomings of existing or future screening systems and performance metrics. In addition, our results could also potentially inform the community about building better automated methods, especially those that benefit from expert knowledge or inventories designed to diagnose mental health conditions.

However, we are aware of the potential harm of automatic screening tools. The mental well-being of individuals is a delicate personal matter that could potentially be exploited to harm people on public online networks. As researchers working with data from social media platforms, we implemented essential measures to safeguard the privacy of individuals, uphold their ethical rights, and prevent any emotional distress. Our actions align with the recommendations provided by Benton et al. [65] and Ayers et al. [66] on data use, storage, and distribution. To guarantee the privacy of the data, we have signed and delivered the confidentiality agreements provided by the data distributors of each experimental framework. All evaluations were performed using data that had been de-identified. Modifications were made to edit or obscure the metadata, ensuring the protection of individuals' privacy in both datasets.

6.3 Theoretical and practical implications

Theoretical implications: Our study highlights that the availability of evidence for different BDI items varies significantly between platforms and symptoms. This suggests that mental health constructs are expressed in different ways in online environments, opening avenues for further research on symptom-specific digital screening. Furthermore, due to the increasing presence of noise, automated systems tend to perform poorly on symptoms with a high online incidence. This finding underscores the need to develop methods to effectively distinguish relevant evidence from irrelevant information.

Our study contributes to understanding how psychometric scales such as the BDI can be integrated into computational frameworks to analyze mental health on social media. This creates opportunities for interdisciplinary research that combines psychology, linguistics, and computer science.

The results highlight the importance of query design in retrieving relevant results for mental health screening. This finding emphasizes the need for advanced natural language processing techniques and the exploration of large language models to improve search methods.

Practical implications: Our findings also highlight the symptoms of depression that can be detected most effectively from user-generated content. This information can guide the development of more accurate and targeted automated screening tools. For example, our study suggests that symptoms with higher incidence levels (e.g., sleep disturbances, indecisiveness) are more readily identifiable on social media. Automated systems can prioritize these symptoms for an initial assessment to improve efficiency and reliability. Researchers who seek to classify individuals based on BDI levels (or related mental health indicators) may benefit from the

insights and methods we propose here, particularly related to the creation of queries or filter profiles to detect excerpts related to BDI within social media data.

The comparative analysis of Reddit and Twitter shows platform-specific variations in symptom representation. This insight could help tailor mental health assessment tools for social media platforms, increasing their usability and accuracy.

Finally, by identifying linguistic and behavioral indicators of depression, health organizations could deploy early warning systems on social media platforms to monitor mental health trends in the population and intervene proactively.

7 Conclusions and future work

In this study, we conducted a comprehensive examination of automated estimation methods of depression assessment inventories, specifically focusing on BDI and using evidence sourced from social media. Our investigation delved into the prevalence of the 21 BDI items within users' feeds, assessing how its frequency of occurrence and other related features impact the efficacy of computer-based tools designed to extract depression symptoms from social media. To measure the occurrence of different depression symptoms, we formulate queries based on BDI items, collect documents from social media collections, and analyze their relevance scores.

Our analysis revealed a consistent occurrence of specific items across various search methods and platforms, indicating a higher prevalence of certain symptoms in these collections. However, abundant documentary evidence on these topics also introduces increased noise. In particular, systems that were designed to measure the severity of depression symptoms from user data exhibited poor performance on BDI items with high incidence. In contrast, items with low incidence were more easily estimated by automated means. Additionally, we observed an inverse correlation between the effectiveness of automatic systems and other features, such as the length of textual descriptions for BDI items.

Our findings open several potential directions for further research. Specifically, we are keen on exploring alternative methods to formulate meaningful queries based on the BDI items. In this work, we followed a simple yet high-precision strategy that could serve as a baseline for future experiments. The generation of succinct and on-topic queries from BDI items represents a challenging task. The quality and specificity of the derived queries directly affect the relevance of the retrieved documents and, subsequently, the effectiveness of any system that mines depression symptoms from the retrieved data. Additionally, an intriguing avenue for exploration involves comparing automatically generated queries with those formulated by human experts, such as psychologists and practitioners. This comparison aims to identify differences and similarities, shedding light on how each strategy could complement the other.

Future research could leverage anonymized electronic health records alongside social media data or other types of patient data (e.g., clinical interview transcripts). This approach would allow us to identify individuals struggling with depression based on diagnostic codes. These patients do not necessarily disclose their condition online, and thus, they could represent a solid complement to our online users. By analyzing textual extracts from this anonymized group, we could

gain valuable insights into the incidence of BDI topics across different data types. In addition, we are interested in understanding the differences between the individuals diagnosed and those who are experiencing symptoms but are not yet seeking help. This would ultimately contribute to developing more comprehensive screening tools to identify at-risk individuals, even if they have not explicitly disclosed their struggles.

An additional area for future research involves exploring the degree to which the frequency of occurrence is related to the significance assigned to a specific item in the questionnaire. While the BDI was initially designed with all items carrying equal weight in assessing an individual's depression severity, a closer examination of the inventory's items raises the question of whether certain items hold more significance than others. For example, imagine an individual who scores higher on the BDI item related to suicide compared to the item that evaluates loss of interest in sex. This may suggest a higher likelihood of depression and a pressing need for intervention to prevent further consequences. Taking into account this perspective, investigating the relationship between the importance of items in estimating depression severity and their frequency and specificity can improve our understanding, guiding the development of automatic assessment tools with a focus on specific items over others.

Our work has also presented a preliminary approach to integrating LLMs in query formulation. However, due to the advanced capabilities of LLMs, these generative systems can be further exploited in future research, e.g., in the context of sophisticated query formulation or search strategies. For example, our work currently uses a straightforward lexical matching approach with queries generated from BDI item titles and responses. This method may overlook other expressions or nuanced linguistic variations of depression symptoms. LLMs could create semantically rich and context-aware descriptions of BDI items. By fine-tuning LLMs with mental health-related corpora or explicitly prompting the models to generate diverse paraphrases for each BDI item, the retrieval process could account for multiple variations in how symptoms are expressed across different user groups and social media platforms. In relation to this, LLMs could support the extraction and analysis of contextual evidence, for example, by analyzing clusters of posts, summarizing trends, and inferring relationships between posts. An LLM could also analyze temporal patterns or subtle references to depression symptoms that emerge across multiple posts, offering a user-level perspective (instead of a post-level analysis). For cross-platform purposes, LLMs fine-tuned on data from multiple platforms could help generalize the methods by identifying platform-specific linguistic patterns. Finally, LLMs could also play a role in dynamic symptom modeling, where we could model relationships between BDI symptoms by analyzing co-occurrence patterns in user-generated content.

Acknowledgements MEA and DEL thank the support obtained from: Agencia Estatal de Investigación (Spain) (PID2022-137061OB-C22; PLEC2021-007662 MCIN/AEI/10.13039/501100011033, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU), Consellería de Cultura, Educación, Formación Profesional e Universidades (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04 and Reference Competitive Group accreditation 2022-2025, ED431C 2022/19) and the European Union (European Regional Development Fund—ERDF).

Author contributions EAR: conceptualization, methodology, investigation, and writing—original draft. MEA: formal analysis, writing, and review. DEL and FC: writing—review, editing, supervision, and project administration.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. The Swiss Government Excellence Scholarships and the Hasler Foundation partly supported this work. This research was also funded by Xunta de Galicia and Ministerio de Ciencia e Innovación (Spain).

Data availability The datasets used in this study are available upon request to the organizers in the following links: CLPsych (https://www.cs.jhu.edu/~mdredze/clpsych-2015-shared-task-evaluation/) and eRisk (https://erisk.irlab.org/).

Declarations

Conflict of interest On behalf of all authors, the corresponding author declares that there is no conflict of interest.

References

1. Skaik, R., & Inkpen, D. (2020). Using social media for mental health surveillance: A review. *ACM Computing Surveys, 53*(6), 1–31.
2. Chancellor, S., & Choudhury, M. D. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine, 3*(1), 43.
3. Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction, 27*(5), 1–53.
4. Saha, K., Seybolt, J., Mattingly, S.M., Aledavood, T., Konjeti, C., Martinez, G. J., Grover, T., Mark, G., & De Choudhury, M. (2021). What life events are disclosed on social media, how, when, and by whom? In *Proceedings of the 2021 CHI conference on human factors in computing systems, CHI '21* Yokohama, Japan.
5. Zarrinkalam, F., Kahani, M., & Bagheri, E. (2018). Mining user interests over active topics on social networks. *Information Processing & Management, 54*(2), 339–357.
6. Khodabakhsh, M., Fani, H., Zarrinkalam, F., & Bagheri, E. (2018). Predicting personal life events from streaming social content. In *Proceedings of the 27th ACM international conference on information and knowledge management, CIKM 2018*, Torino, Italy (pp. 1751–1754).
7. Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: A good place to detect health conditions. *PLoS ONE, 9*(1), 1–11.
8. Correia, R. B., Wood, I. B., Bollen, J., & Rocha, L. M. (2020). Mining social media data for biomedical signals and health-related behavior. *Annual Review of Biomedical Data Science, 3*(1), 433–458.
9. Ramírez-Cifuentes, D., Freire, A., Baeza-Yates, R., Sanz Lamora, N., Álvarez, A., González-Rodríguez, A., Lozano Rochel, M., Llobet Vives, R., Velazquez, D. A., Gonfaus, J. M., & Gonzàlez, J. (2021). Characterization of anorexia nervosa on social media: Textual, visual, relational, behavioral, and demographical analysis. *Journal of Medical Internet Research, 23*(7), 25925.

10. Ríssola, E. A., Losada, D. E., & Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare, 2*(2), 17–11731.
11. Shah, A. M., Yan, X., Tariq, S., & Ali, M. (2021). What patients like or dislike in physicians: Analyzing drivers of patient satisfaction and dissatisfaction using a digital topic modeling approach. *Information Processing & Management, 58*(3), 102516.
12. Cheng, P. G. F., Ramos, R. M., Bitsch, J. Á., Jonas, S. M., Ix, T., See, P. L. Q., & Wehrle, K. (2016). Psychologist in a pocket: Lexicon development and content validation of a mobile-based app for depression screening. *JMIR mHealth and uHealth, 4*(3), 88.
13. Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*(1), 547–577.
14. Coppersmith, G., Harman, C., & Dredze, M. (2014). Measuring post traumatic stress disorder in twitter. In *Proceedings of the 8th international conference on weblogs and social media, ICWSM 2014*, Ann Arbor, USA.
15. Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. In *Proceedings of the ninth international conference on web and social media, ICWSM 2015*, Oxford, UK (pp. 31–40).
16. Ahmad, P. N., Yuanchao, L., Aurangzeb, K., Anwar, M. S., & Haq, Q.M.u. (2024). Semantic web-based propaganda text detection from social media using meta-learning. *Service Oriented Computing and Applications*
17. Fernández-Pichel, M., Aragón, M. E., Saborido-Patiño, J., & Losada, D. E. (2023). Personality trait analysis during the covid-19 pandemic: A comparative study on social media. *Journal of Intelligent Information System, 62*(1), 117–142.
18. Kazai, G., Thomas, P., Craswell, N. (2019). The emotion profile of web search. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, SIGIR'19*, Paris, France (pp. 1097–1100).
19. Milton, A., & Pera, M. S. (2020). What snippets feel: Depression, search, and snippets. In *Proceedings of the joint conference of the information retrieval communities in Europe (CIRCLE 2020)*, Samatan, France.
20. Losada, D. E., Crestani, F. A., & Parapar, J. (2017). erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. In *Conference and labs of the evaluation forum*.
21. Losada, D.E., Crestani, F., & Parapar, J. (2018). Overview of erisk: Early risk prediction on the internet. In *Experimental IR meets multilinguality, multimodality, and interaction—9th international conference of the CLEF Association, CLEF 2018*, Avignon, France (pp. 343–361).
22. Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR meets multilinguality, multimodality, and interaction—10th international conference of the CLEF Association, CLEF 2019*, Lugano, Switzerland (pp. 340–357).
23. Losada, D. E., Crestani, F., & Parapar, J. (2020). Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR meets multilinguality, multimodality, and interaction—11th international conference of the CLEF Association, CLEF 2020*, Thessaloniki, Greece (pp. 272–287).
24. Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2021). Overview of erisk 2021: Early risk prediction on the internet. In K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, & N. Ferro (Eds.), *Experimental IR meets multilinguality, multimodality, and interaction* (pp. 324–344). Springer.
25. Crestani, F., Losada, D. E., & Parapar, J. (2022). Early detection of mental health disorders by social media monitoring: The first five years of the eRisk project. Studies in Computational Intelligence. Springer.
26. Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2023). Overview of erisk 2023: Early risk prediction on the internet. In *Experimental IR meets multilinguality, multimodality, and interaction: 14th international conference of the CLEF Association, CLEF 2023*, Thessaloniki, Greece, September 18–21, 2023, Proceedings (pp. 294–315). Springer.
27. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). 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* (pp. 31–39).
28. Masood, R. (2019). Adapting models for the case of early risk prediction on the internet. In *Advances in information retrieval—41st European Conference on IR Research, ECIR 2019*, Cologne, Germany (pp. 353–358).

29. Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the beck depression inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*(1), 77–100.
30. Urbina, S., & Anastasi, A. (1997). *Psychological testing* (7th ed.). Prentice Hall.
31. Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the seventh international conference on weblogs and social media, ICWSM 2013*, Cambridge, USA.
32. Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences, 18*(Supplement C), 43–49.
33. Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An Inventory for Measuring Depression. *Archives of General Psychiatry, 4*(6), 561–571.
34. Sik, D., Rakovics, M., Buda, J., & Németh, R. (2023). The impact of depression forums on illness narratives: A comprehensive NLP analysis of socialization in e-mental health communities. *Journal of Computational Social Science*.
35. Radloff, L. S. (1977). The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385–401.
36. Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with twitter data. *Scientific Reports, 7*(1), 13006.
37. Chu, M. Y. A., Chan, L. S. H., Chang, S. S.-Y., Tiwari, A., Yuk, H., & So, M. K. P. (2024). Applications of bayesian networks in assessing the effects of family resilience on caregiver behavioral problems, depressive symptoms, and burdens. *Journal of Computational Social Science 7*(2), 1275–1303.
38. Skaik, R. S., & Inkpen, D. (2022). Predicting depression in Canada by automatic filling of Beck's depression inventory questionnaire. *IEEE Access, 10*, 102033–102047.
39. Kang, M., Oh, S., Oh, K., Kang, S., & Lee, Y. (2021). The deep learning method for predict beck's depression inventory score using EEG. In *2021 International conference on information and communication technology convergence (ICTC)* (pp. 490–493).
40. Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., & Ungar, L. (2014). Towards assessing changes in degree of depression through facebook. In *Workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*.
41. Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe, 7*, 7–28.
42. Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD workshop on healthcare informatics*
43. Association, A. P. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). American Psychiatric Publishing.
44. Gaur, M., Kursuncu, U., Alambo, A., Sheth, A., Daniulaityte, R., Thirunarayan, K., & Pathak, J. (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* (pp. 753–762).
45. Yazdavar, A.H., Al-Olimat, H.S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., & Sheth, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social network analysis and mining. IEEE/ACM international conference on advances in social network analysis and mining 2017* (pp. 1191–1198).
46. Kroenke, K., Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2010). The patient health questionnaire somatic, anxiety, and depressive symptom scales: A systematic review. *General Hospital Psychiatry, 32*(4), 345–359.
47. Guo, Z., Ding, N., Zhai, M., Zhang, Z., & Li, Z. (2023). Leveraging domain knowledge to improve depression detection on Chinese social media. *IEEE Transactions on Computational Social Systems, 10*(4), 1528–1536.
48. Ghosh, S., & Anwar, T. (2021). Depression intensity estimation via social media: A deep learning approach. *IEEE Transactions on Computational Social Systems, 8*(6), 1465–1474.
49. Ezerceli, Ö., & Dehkharghani, R. (2024). Mental disorder and suicidal ideation detection from social media using deep neural networks. *Journal of Computational Social Science*, *7*, 2277–2307.
50. Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2022). What does your bio say? inferring twitter users' depression status from multimodal profile information using deep learning. *IEEE Transactions on Computational Social Systems, 9*(5), 1484–1494.

51. Sadasivuni, S.T., & Zhang, Y. (2020). A new method for discovering daily depression from tweets to monitor peoples depression status. In *2020 IEEE international conference on humanized computing and communication with artificial intelligence (HCCAI)* (pp. 47–50).
52. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2024). Detecting mental disorder on social media: A ChatGPT-augmented explainable approach.
53. Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International workshop on semantic evaluation (SemEval-2017)* (pp. 747–754). Association for Computational Linguistics, Vancouver, Canada.
54. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of The Third Text REtrieval conference*, TREC 1994, Gaithersburg, Maryland, USA (pp. 109–126).
55. Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, Melbourne, Australia (pp. 275–281).
56. Croft, B., Metzler, D., & Strohman, T. (2009). *Search engines: Information retrieval in practice* (1st ed.). Addison-Wesley Publishing Company.
57. Mitra, B., & Craswell, N. (2018). An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval, 13*(1), 1–126.
58. Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI conference on human factors in computing systems. CHI '07* (pp. 417–420).
59. Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, Baltimore, USA
60. Yang, P., Fang, H., & Lin, J. (2018). Anserini: Reproducible ranking baselines using lucene. *Journal of Data and Information Quality 10*(4).
61. Fuhr, N. (2018). Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum, 51*(3), 32–41.
62. Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. The Guilford Press.
63. Cheng, S., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., & Wang, L. (2023). Prompting gpt-3 to be reliable. In *International conference on learning representations (ICLR 23)*.
64. Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of twitter research ethics. *Social Media + Society 4*(1)
65. Benton, A., Coppersmith, G., & Dredze, M. (2017). Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, EthNLP@EACL, Valencia, Spain, April 4, 2017 (pp. 94–102).
66. Ayers, J. W., Caputi, T. L., Nebeker, C., & Dredze, M. (2018). Don't quote me: Reverse identification of research participants in social media studies. *npj Digital Medicine, 1*(1), 30.

Authors and Affiliations

Esteban A. Ríssola[1] · Mario Ezra Aragón[2] · David E. Losada[2] · Fabio Crestani[3]

✉ Mario Ezra Aragón
ezra.aragon@usc.es

Esteban A. Ríssola
earissola@unlu.edu.ar

David E. Losada
david.losada@usc.es

Fabio Crestani
fabio.crestani@usi.ch

[1] Departamento de Ciencias Básicas, Universidad Nacional de Luján (UNLu), 6700 Buenos Aires, Argentina

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

[3] Faculty of Informatics, Università della Svizzera italiana (USI), 6900 Lugano, Switzerland