Online Social Networks and Media 50 (2025) 100339

Contents lists available at ScienceDirect

Online Social Networks and Media

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

How does depression talk on social media? Modeling depression language with relevance-based statistical language models

Eliseo Bao *,[1], Anxo Perez [1], David Otero [1], Javier Parapar [1]

IRLab, CITIC, Universidade da Coruña, A Coruña, Spain

ARTICLE INFO

Keywords:
Mental health
Depression
Language modeling
Natural language processing
Text mining
Social media
User risk assessment
Clinical markers
Linguistic patterns
Psycholinguistics

ABSTRACT

Many individuals with mental health problems turn to the internet and social media for information and support. The text generated on these platforms serves as a valuable resource for identifying mental health risks, driving interdisciplinary research to develop models for mental health analysis and prediction. In this paper, we model depression-related language using relevance-based statistical language models to create lexicons that characterize linguistic patterns associated with depression. We also propose a ranking method that leverages these lexicons to prioritize users exhibiting stronger signs of depressive language on social media. Our models integrate clinical markers from established depression questionnaires, particularly the Beck Depression Inventory-II (BDI-II), enhancing explainability, generalization, and performance. Experiments across multiple social media datasets show that incorporating clinical knowledge improves user ranking and generalizes effectively across platforms. Additionally, we refine existing depression lexicons by applying weights estimated from our models, achieving better performance in generating depression-related queries. A comparative analysis of our models highlights differences in language use between control users and those with depression, aligning with prior psycholinguistic findings. This work advances the understanding of depression-related language through statistical modeling, paving the way for scalable social media interventions to identify at-risk individuals.

1. Introduction

Mental health is essential for overall well-being, yet over 20% of adults experience mental disorders [1,2]. Among these, depressive disorder is a leading condition, characterized by persistent low mood or loss of interest in activities, affecting approximately 332 million people globally [3]. Early intervention is critical, particularly for young individuals [4]. However, many at-risk individuals do not seek care due to stigma: over 60% of those with depression avoid treatment for this reason [5,6]. Financial constraints, lack of insurance, and limited access to mental health services further hinder care, especially in underserved regions [7].

To address inattention challenges, computational researchers increasingly leverage social media content to detect signs of mental health disorders, aiming to mitigate their societal impact [8,9]. Social media provides valuable insights into users' mental states, as individuals often express thoughts and emotions more openly due to perceived privacy and anonymity [10–12]. Textual analysis of writing styles has been effective in identifying conditions like depression [13], offering a data-rich alternative to traditional therapeutic settings.[2] Advances in computational linguistics, supported by curated benchmarks [14,15], have improved depression detection models. However, these models are designed to complement, not replace, mental health professionals, requiring rigorous validation and clinician oversight [16]. Trust mechanisms, such as validated clinical questionnaires, are critical for professional confidence [17]. Tools like the Beck Depression Inventory-II (BDI-II) [18] and the 9-item Patient Health Questionnaire (PHQ-9) [19] assess symptoms such as sadness, irritability, and sleep disturbances, enhancing detection methods' explainability, generalization, and performance [20].

Research in this area has progressed from traditional methods using engineered features like word counts, posting activity, and emotion levels [21,22] to state-of-the-art Large Language Models (LLMs) [23–25]. Transformer-based approaches now serve as classifiers to detect users at risk of depression and related disorders in online settings [26,27]. Although models such as BERT and RoBERTa have shown strong predictive performance, their lack of interpretability and sensitivity to

* Corresponding author.
E-mail addresses: eliseo.bao@udc.es (E. Bao), anxo.pvila@udc.es (A. Perez), david.otero.freijeiro@udc.es (D. Otero), javier.parapar@udc.es (J. Parapar).

[1] These authors contributed equally to this work.

[2] https://wearesocial.com/uk/blog/2022/01/digital-2022/

https://doi.org/10.1016/j.osnem.2025.100339
Received 3 July 2025; Received in revised form 19 September 2025; Accepted 17 October 2025
Available online 22 October 2025

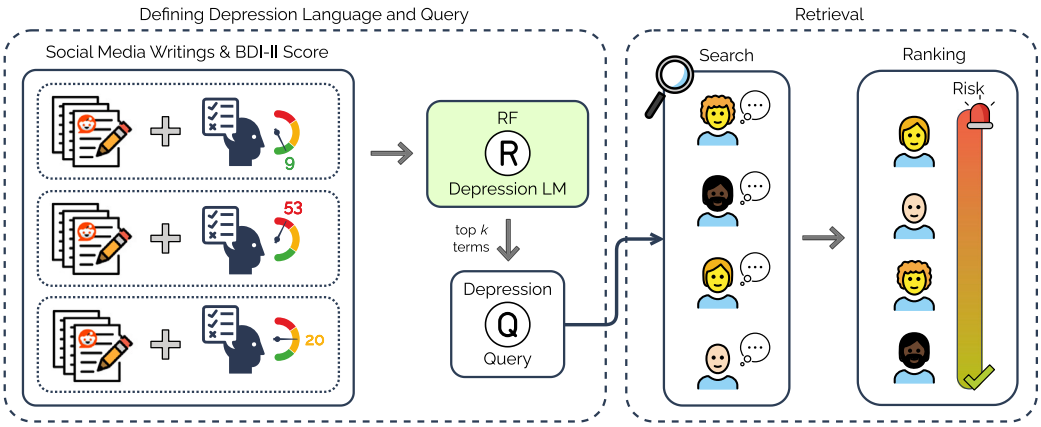


Fig. 1. Overall pipeline of our proposal.

dataset and platform shifts can limit their practical applicability. For example, Novikova and Shkaruta [28] show that BERT-based classifiers often struggle to generalize across datasets and fail to identify clinically important markers such as suicidal ideation. This highlights the risk of relying on opaque models, particularly in domains where understanding how a decision is made is as important as the decision itself. In contrast, Relevance-Based Statistical Language Models, commonly used in the *ad hoc* retrieval task for *pseudo-relevance feedback* (PRF), remain largely unexplored for mental health applications. These models estimate term distributions over users' language, making it possible to observe which expressions contribute to predictions. This interpretability is especially valuable in clinical or public health settings, where transparency and human oversight are essential.

As shown in Fig. 1, we propose adapting Relevance-Based Statistical Language Models by replacing the *query likelihood* with users' BDI-II scores. This adjustment gives greater weight to vocabulary used by individuals with higher BDI-II scores. Users' entire writing activity is treated as documents, enabling the estimation of term weights for the full vocabulary. This process generates what we call *depression languages*, i.e. lists of term-weight pairs reflecting the language patterns of individuals with depression. These *depression languages* are versatile, usable in various tasks, and form the basis of our pipeline to select terms and construct queries for a classical retrieval system, optionally incorporating validated depression lexicons [8,29]. The system ranks users by the similarity of their writing to the constructed depression query, identifying those at higher risk of depression. Unlike most studies, which treat depression detection as a binary classification task [30,31], our approach provides a ranked estimation. To contextualize our probabilistic framework, we also benchmark it against a strong baseline transformer-based ranker. Technical details of this method are discussed in Section 3.

A notable advantage of the proposed systems is their computational efficiency across training, deployment, and inference. Unlike deep learning approaches, Relevance-Based Statistical Language Models require minimal computational resources, making them feasible to run on inexpensive hardware. This accessibility promotes the democratization of such systems. Additionally, these probabilistic models estimate a probability distribution over language, offering insights into how individuals with depression express themselves. The measurable term weights in the language of those with depression enhance transparency and build trust with clinicians, an essential factor in the adoption of these systems [32].

As a result of our proposal, we have: (i) developed a method to compute vocabularies that model the language of depression, capturing patterns in how individuals with this condition communicate; and (ii) created a pipeline that leverages these vocabularies to detect at-risk users and rank them by their estimated depression levels. This ranking capability is particularly valuable during triage, enabling professionals to prioritize those at higher risk. Such a system is especially beneficial in population screening, identifying seemingly healthy individuals who may be at increased risk of developing depression or related conditions.

We have open-sourced the complete code for this project, including the depression language models, in the following repository: https://gitlab.irlab.org/eliseo.bao/rf-models-depression-language. Through this work, we aim to explore the following research questions:

- **RQ1** Are relevance-based statistical language models effective for modeling depression language?
- **RQ2** Can the model be generalized to other social media platforms?
- **RQ3** Can relevance-based statistical language models enhance existing lexicons?
- **RQ4** What insights do these vocabularies provide?

2. Related work

Research on depression predates the Internet and has traditionally relied on user studies and questionnaires, leading to widely accepted tools such as Beck's Depression Inventory (BDI) [33] and the CES-D Scale [34]. The BDI includes 21 questions assessing mental and physiological states, while the CES-D Scale uses 20 questions to address conditions like guilt and sleep disturbances. These scales typically assign scores based on symptom severity, determining depression levels. Similarly, the Diagnostic and Statistical Manual of Mental Disorders (DSM) [35] outlines 9 indicators, such as depressed mood and diminished interest in activities, which clinicians use to diagnose depression based on symptom duration. Early research also linked mental health conditions to linguistic features. Beck [36] developed cognitive therapy emphasizing the use of negatively valenced[3] words. Pyszczynski et al. [37] studied first-person pronouns and negative expectations, while Al-Mosaiwi and Johnstone [38] identified absolutist language (e.g., "always", "nothing") as a marker for depression, anxiety, and suicidal ideation. Losada and Crestani [39] created a test collection for evaluating language-based depression detection methods. Subsequent studies have validated links between linguistic features and mental states [40,41]. With the advent of social media, researchers now analyze user-generated content to study mental health [42,43]. For instance, Coppersmith et al. [44] examined linguistic patterns in tweets from individuals self-reporting mental health conditions like ADHD and depression, identifying communication markers unique to these conditions.

With the rapid growth of online text data and the sensitivity of mental health issues, manual analysis and timely psychiatric intervention have become impractical at scale. To address this, Natural Language

[3] https://dictionary.apa.org/valence

Processing (NLP) and text mining approaches have been employed for automated analysis of social media data. Recent advances in deep learning enable automatic capture of latent semantic features, reducing reliance on manual feature engineering. Transformer-based pre-trained language models (PLMs) like BERT [45] and RoBERTa [46] have proven effective in detecting mental health issues such as depression and suicidal ideation. Specialized models, such as Mental-BERT [26] and DisorBERT [47], are tailored for analyzing social media data in mental health contexts. While these approaches fall within the domain of machine learning, probabilistic methods from Information Retrieval have also been applied to sentiment analysis [48–50].

Most of the work reviewed thus far focuses on determining whether social media users are positive for a mental health condition. However, there is growing interest in approaches that estimate a depression score [51]. The Early Risk Prediction on the Internet (eRisk)[4] framework introduced a ranking-based case study for early depression detection in its 2022 edition [52]. In this task, systems assign a score representing a user's estimated risk level, enabling the computation of ranking-based metrics. During training, participants had access to users' complete writing histories. In the test phase, user writings were released incrementally, requiring estimations after 1, 100, 500, and 1000 writings. Transformer-based systems #0 and #1 from the BLUE team [53] demonstrated the most consistent performance across varying data availability. This framework is particularly relevant as it forms the basis for our experiments, detailed in Section 4.

3. Task and method

This section is organized as follows: Section 3.1 defines and formalizes the task we address. Section 3.2 presents our proposed solution to this task. Finally, Section 3.3 details the context of our method's origins and outlines the formulation of the models used in our research.

3.1. Task

The primary goal of this work is to model the language associated with depression by evaluating the weights of words in the vocabulary used by individuals experiencing this condition. This produces what we term *depression language*: lists of words paired with their respective weights. To achieve this, we employ relevance-based statistical language models, originally developed for ad-hoc search in information retrieval (IR). These models estimate $p(w|R)$, the probability of a word w given a concept of relevance R. In our case, relevance is defined as depression, leading to the computation of $p(w|\text{Dep})$ for each word in the vocabulary. Formally, given a vocabulary $V = \{w_1, w_2, \ldots, w_n\}$, our task is to estimate $p(w|\text{Dep})$ for each $w \in V$, where Dep represents the condition of depression. This probability quantifies the relevance of each word to depression.

3.2. Our proposal

Our proposal adapts Section 3.3 to model the language associated with depressive disorders. A detailed explanation of these models and their application in Information Retrieval (IR) is provided in the following section. At its core, the approach leverages term weighting in pseudo-relevance based statistical language models to expand queries and improve document ranking. These models traditionally calculate term weights based on the *query likelihood*, which represents the probability of a document being relevant to a query. In our adaptation, each document corresponds to a user's entire writing history on a social media platform (see Section 4.1), accompanied by the user's Beck Depression Inventory-II (BDI-II) score [18]. Since higher BDI-II scores reflect more severe depression, we propose replacing the *query likelihood* factor with the BDI-II score. This score acts as an estimator of a user's influence in the language modeling process.

As detailed in Section 3.1, this work aims to model depression language using relevance-based statistical language models. To achieve this, we first construct an indexed collection where each document represents a user's complete writing history, along with their BDI-II depression score. We then adapt the formulations of RM (Section 3.3.1), DMM (Section 3.3.2), and MEDMM (Section 3.3.3) to incorporate the BDI-II score as a weight reflecting the influence of a user's language on the depression model. Using these adapted models, we estimate term-weight pairs for the vocabulary, considering various relevance set sizes. The resulting *depression languages* consist of vocabulary terms, each assigned a weight. Higher weights indicate terms that are more *important* in modeling depression language.

By integrating BDI-II scores, a clinically validated measure of depressive symptomatology, into the language modeling process, we ensure that the derived depression language is not inferred through assumed lexical markers but is instead weighted by users' psychological states. Such an approach is consistent with prior psychological research demonstrating that everyday language use reflects personality traits, social context and mental health conditions [54]. This underscores the value of using BDI-II–informed user text to capture patterns in depression-related discourse.

It is important to note that our proposal is focused to textual features. The eRisk datasets used in this work do not provide interactional or relational metadata (e.g., replies, likes, network connections), and incorporating such features would also hinder comparability with prior eRisk baselines, which are defined solely on textual content. While this narrows the scope to linguistic signals, our framework remains scalable and could be readily integrated with future approaches that include relational or behavioral features to capture a more holistic view of depression in online settings.

3.3. Relevance-based statistical language models

Ad hoc retrieval is a fundamental task in Information Retrieval (IR) [55], involving the search for documents relevant to a user's information need, typically expressed as a *query*. However, users often struggle to accurately articulate their needs in query form. Query expansion (QE), which enhances a query by adding new terms, is a proven method for improving retrieval effectiveness [56]. Among QE techniques, *relevance feedback* is one of the most effective [57,58]. While *real relevance feedback* relies on users identifying relevant documents, this process is often expensive or infeasible. Consequently, *pseudo-relevance feedback* (PRF), which assumes the top-k documents retrieved are relevant, has become predominant [56]. PRF extracts terms and weights from these documents to expand the original query. Statistical language models [59,60] are among the most robust PRF techniques, combining strong theoretical foundations with empirical success [61]. This work focuses on three such models: The Relevance-Based Statistical Language Models (RM) proposed by Lavrenko and Croft [62], the Divergence Minimization Model (DMM) introduced by Zhai and Lafferty [63] and Maximum-Entropy Divergence Minimization Model (MEDMM), an expansion of DMM proposed by Lv and Zhai [64]. We detail these models in Sections 3.3.1 to 3.3.3, respectively.

3.3.1. RM

Relevance-Based Statistical Language Models (commonly referred to as Relevance Models or RM) were developed to explicitly integrate the concept of *relevance*, a core principle of probabilistic models, into language modeling [62]. In an RM, the original *query* is treated as a sample of words drawn from the relevance model itself (R). When additional words are selected from R, those with the highest estimated probabilities are chosen, based on the observed distribution of words. Terms in the collection's lexicon are then ranked by their

[4] https://erisk.irlab.org/

estimated probability, calculated under the independent and identically distributed (i.i.d.) sampling assumption, as shown in Eq. (1).

$$p(w|R) = \sum_{D \in C} p(w|D)\, p(D) \prod_{i=1}^{n} p\left(q_i|D\right) = \sum_{D \in C} p(w|D) \prod_{i=1}^{n} p\left(q_i|D\right) \quad (1)$$

In Eq. (1), the document prior $p(d)$ is generally assumed to be uniform. The term $\prod_{i=1}^{n} p(q_i|D)$ represents the *query likelihood* under the document model, while $p(w|D)$ denotes the term probability given a document D. Both probabilities are typically estimated using smoothing techniques to handle sparse data effectively.

Adapted model. In the proposal section (see Section 3.2), we explained that we replace the *query likelihood* factor with BDI-II scores. The modified RM formulation is shown in Eq. (2), where $p(w|R)$ represents the probability of the word w being relevant to the depression language R. The summation $\sum_{d \in C}$ spans all documents d in the collection C, where each document corresponds to a specific user's writings. The term $p(w|D)$ denotes the probability of w appearing in the user's document D, and BDI-II(D) is the BDI-II score associated with the user. This score adjusts the document's weight, emphasizing contributions from users with higher BDI-II scores, under the assumption that their language is more indicative of depression-related patterns.

$$p(w|R) = \sum_{d \epsilon C} \cdot p(w|D) \cdot \text{BDI-II}(D) \quad (2)$$

3.3.2. DMM

The Divergence Minimization Model (DMM) [63] is a PRF technique that assumes the feedback model θ_F should closely resemble the language model of the relevant documents F while diverging significantly from the background model. The model is computed according to Eq. (3).

$$p(w|\theta_F) \propto \exp\left(\frac{1}{1-\lambda}\frac{1}{|F|}\sum_{d \in F} \log p(w|\theta_d) - \frac{\lambda}{1-\lambda}\log p(w|\theta_C)\right) \quad (3)$$

where $p(w|\theta_d)$ is typically computed using additive smoothing as recommended by Hazimeh and Zhai [65] (Eq. (4)).

$$p(w|\theta_d) = \frac{tf(w,d) + \gamma}{|d| + \gamma \cdot |V|} \quad (4)$$

This model includes a parameter λ to control the influence of the collection language model and a parameter γ to control the smoothed document model. Since no *query likelihood* factor is included in the formulation, the model can be applied to our task without requiring further adaptation.

3.3.3. MEDMM

The Maximum-Entropy Divergence Minimization Model (MEDMM) [64] is an RF technique derived from DMM [63]. While built on the same principles as DMM, MEDMM addresses some of its limitations. The model is formulated as an optimization problem, and applying the Lagrange Multiplier method yields the analytical solution presented in Eq. (5).

$$p(w|\theta_F) \propto \exp\left(\frac{1}{\beta}\sum_{d \in F} \alpha_d \log p(w|\theta_d) - \frac{\lambda}{\beta}\log p(w|\theta_C)\right) \quad (5)$$

where $p(w|\theta_d)$ and $p(w|\theta_C)$ are computed as described in Section 3.3.2. This model introduces two parameters: λ, which controls the Inverse Document Frequency (IDF) effect by assigning greater importance to terms that occur less frequently in the collection (i.e., terms with higher IDF [66]), and β, which regulates the entropy of the feedback language model. Unlike DMM, which assigns equal weights to all feedback documents (setting $\alpha_d = \frac{1}{|F|}$), MEDMM assigns varying weights to feedback documents based on the posterior probability of the document language model (Eq. (6)).

$$\alpha_d = p(\theta_d|q) = \frac{p(q|\theta_d)}{\sum_{d' \in F} p(q|\theta_{d'})} = \frac{\prod_{w \in q} p(w|\theta_d)}{\sum_{d' \in F} \prod_{w' \in q} p(w'|\theta_{d'})} \quad (6)$$

Adapted model. After adapting MEDMM to incorporate the BDI-II score, the resulting formulation is shown in Eq. (7). In this equation, α_d represents the weight assigned to document d, corresponding to the writings of a specific user. The numerator, $\prod_{w \in q} \text{BDI-II}(d)$, incorporates the BDI-II score of the user associated with d, reflecting the extent to which the user's writings influence the depression language.

$$\alpha_d = \frac{\prod_{w \in q} \text{BDI-II}(d)}{\sum_{d' \in F} \prod_{w' \in q} p(w'|\theta_{d'})} \quad (7)$$

4. Experiments and results

The proposed experiments aim to showcase practical use cases of the depression languages we developed, demonstrating their applicability to real-world scenarios and evaluating their quality. While **RQ4** focuses on an analytical study, **RQ1** to **RQ3** are evaluated in ranking-based setups. In these setups, depression languages are integrated into a system that processes social media posts from a group of users, ranking them based on their estimated risk of depression. Formally, let $U = \{u_i\}$ represent a set of n users, where each user u_i is characterized by their published social media writings $u_i = \{w_{i,j}\}$, with $w_{i,j}$ denoting the jth post of user u_i. Here, j reflects the chronological order of posts, such that if $w_{i,j}$ was published before $w_{i,k}$, then $j < k$ $(j,k \geq 1)$. Given a subset of m writings per user $(u_i = \{w_{i,j} \mid j \leq m\})$, the objective is to generate a ranking of users U. For two users at positions p and q in the ranking, if $p < q$ (lower position values indicate higher ranks), the user at p is estimated to have a higher risk of depression than the user at q. This setup mirrors a classic retrieval task, where documents are ranked in response to a query. In this context, we derive *depression queries* by selecting terms from the depression language and retrieving the most relevant documents (users) based on these queries. The resulting ranking prioritizes users whose language most closely aligns with the depression query, positioning them at the top.

We organize the section into six subsections. Section 4.1 describes the datasets used in our experiments, including their sources and characteristics. Section 4.2 outlines the evaluation metrics and framework. Sections 4.3, 4.4, 4.5, and 4.6 present the experiments conducted to address the defined research questions, along with an analysis of the results and findings for each case.

4.1. Datasets

All the data used in our methods and experiments are sourced from two well-established benchmarks in the mental health domain: Early Risk Prediction on the Internet (eRisk)[5] and Computational Linguistics and Clinical Psychology (CLPsych).[6] The eRisk competition focuses on developing methods for early detection of mental health risks such as depression, anorexia, and self-harm, while the CLPsych initiative explores the intersection of NLP and clinical psychology to advance mental health understanding and treatment. Both workshops provide large-scale datasets derived from online forums and social media platforms, simulating real-world scenarios. eRisk datasets are sourced from Reddit, where users discuss their experiences, while CLPsych datasets are derived from Twitter. Leveraging these benchmarks, we used two types of datasets:

(i) The **BDI-II dataset**, sourced from the eRisk *Measuring the Severity of the Signs of Depression* task, includes users' writings and their BDI-II scores, ranging from 0 to 63. The BDI-II score, based on responses to questions about depression symptoms, reflects the severity of a user's depression, with higher scores indicating greater risk. Users are categorized into four clinical groups: minimal, mild, moderate, and severe depression. We utilized datasets from the 2019 [67], 2020 [68], and 2021 [69] editions of this task. Table 1 summarizes the number of

[5] https://erisk.irlab.org/
[6] https://clpsych.org/

Table 1
Number of users in the BDI-II dataset for each edition, categorized by their depression group.

BDI-II score	Depression group	2019	2020	2021	Total
0–13	minimal	7	24	10	41
14–19	mild	1	12	10	23
20–28	moderate	4	14	26	44
29–63	severe	8	20	34	62
Total		20	70	80	170

Table 2
Statistics on used collections.

	BDI-II dataset	Binary dataset			
	Reddit	Reddit		Twitter	
	Total	Train	Test	Train	Test
Depressed users	–	214	98	327	150
Control users	–	1493	1302	572	300
Total # of users	170	1707	1400	899	450
Total # of writings	78740	1 075 741	899 149	1 993 154	1208113
Avg. writings per user	463	631	642	2217	2684
Min. # of writings	16	9	6	1	1
Max. # of writings	1510	2000	2003	3000	3000

users in each BDI-II score group across these years, totaling 170 users across all editions and categories.

(*ii*) The **Binary Depression datasets**, consisting of binary classifications indicating whether users are diagnosed with depression, along with their writing histories. These datasets include training and test users from eRisk (2022) and CLPsych (2015). Incorporating data from both Reddit and Twitter enables us to evaluate the robustness and generalizability of our methods across different platforms, addressing our research question **RQ2**: *Can the Model be Generalized to Other Social Media Platforms?* This diversity ensures that our findings are not platform-specific and can be applied broadly to online environments where users express mental health concerns. Table 2 presents detailed statistics for all datasets utilized in this study, including the number of depressed users, the total number of writings, and the average number of writings per user.

4.2. Evaluation

For evaluation, we follow the framework established in the 2022 edition of the eRisk workshop for the ranking-based evaluation of the early detection of depression task [52]. We report the two standard ranking metrics included for this framework: Precision@k (P@k), which measures the proportion of relevant users among the top-k ranked positions, and Normalized Discounted Cumulative Gain (NDCG), which accounts for the position of relevant users in the ranking, rewarding higher placements and penalizing lower ones [70]. Full definitions of these metrics are provided in the eRisk overview [52].

4.3. RQ1: *Are Relevance-Based Statistical Language Models Effective for Modeling Depression Language?*

In Section 3, we described our formulation of relevance-based statistical language models incorporating clinical information using the BDI-II score. In this experiment, we compute relevance feedback language models (i.e., weight distributions over the vocabulary) to model depression language using the BDI-II datasets. By leveraging user writings alongside their BDI-II scores, these weight distributions aim to capture the linguistic characteristics of individuals experiencing depression. To evaluate the effectiveness of this modeling, we test whether the term weights estimated by these models can accurately rank users based on their estimated risk of depression. Specifically, we generate depression-related queries using terms from these models and apply them in a classic retrieval scenario, where users correspond to documents containing their writings. The rankings are computed using the Reddit binary depression datasets from eRisk, as outlined in Section 4.1.

The first step involves identifying the optimal number of query terms, the optimal relevance set size, and the retrieval similarity configuration. These variables, detailed in Table 3, are treated as hyperparameters. Query terms are selected as the top e terms from the RMs. For the relevance set size, users are sorted by their BDI-II scores (higher scores indicate a greater presence of depressive symptoms), and different numbers of users are added to the relevance set. For retrieval configurations, we tested Jelinek-Mercer and Dirichlet relevance models with various smoothing parameters. Hyperparameter optimization is conducted on the training split of the Reddit binary depression dataset from eRisk 2022. During this phase, we compute relevance feedback language models for all hyperparameter combinations and evaluate the resulting rankings using the proposed offline metrics to identify the best-performing configuration.

After selecting the optimal hyperparameters, we evaluated the final ranking methods using the eRisk evaluation framework. This framework simulates four early detection scenarios, processing 1, 100, 500, and 1000 writings per user to compute the final ranking. Results are presented in Table 4, comparing non-boosted and boosted methods. For query term selection, the top e terms from the RMs are used. Each term's associated weight determines whether the terms are boosted (boosted methods) or not (non-boosted methods) during retrieval. We also report three competitive baselines from eRisk: the best-performing run, the 90th percentile, and the average of all participant runs in the eRisk 2022 *Early Detection of Depression* task [52], which featured a total of 62 methods. Many eRisk submissions leveraged transformer-based architectures (e.g., BERT/RoBERTa variants), providing a relevant point of reference for our results. In addition, we include a transformer baseline using a cross-encoder: the pre-trained `cross-encoder/ms-marco-MiniLM-L6-v2`,[7] a BERT-style MiniLM model fine-tuned for query–document relevance. We adapt this transformer to our ranking task by computing pairwise relevance scores between the depression query and each user's concatenated writings.

Our experimental results show that both boosted and non-boosted RMs effectively model depression language, with several methods outperforming competitive baselines from the eRisk 2022 task. For instance, with 100 writings per user, the non-boosted RM1 and MEDMM methods achieved a perfect P@10 of 1.00, surpassing the best baseline scores of 0.90 and 0.93, respectively. Boosted models also performed well, with the boosted MEDMM model achieving 0.90 for P@10 and 0.91 for *NDCG*@10, closely matching the top baseline scores. As the number of writings per user increased to 500 and 1000, both boosted and non-boosted models maintained strong performance. For example, with 500 writings, the non-boosted RM1 model achieved an *NDCG*@100 score of 0.66, outperforming the 90th percentile baseline score of 0.61. In the strict setting of processing only one writing per user, non-boosted DMM and MEDMM models achieved P@10 scores of 0.70, exceeding the average baseline score of 0.32. These models also performed well for *NDCG*@10, with MEDMM reaching 0.76 compared to the average baseline of 0.34. However, processing a single writing per user often provides insufficient information to assess depressive states.

The transformer-based cross-encoder baseline, which was adapted to rank users by computing pairwise relevance scores between the depression query and each user's concatenated writings, performed reasonably well in the 1-writing scenario, achieving a P@10 of 0.60 and an *NDCG*@10 of 0.72, outperforming the average eRisk runs and even approaching the 90th percentile of the eRisk results. However, its performance quickly plateaued as more writings were added, with scores remaining flat or even declining across the 100, 500, and 1000 writing conditions. This is largely due to the model's 512-token input

[7] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

Table 3
Hyperparameters and the different values considered for optimization.

Hyperparameter	Swept values
Relevance Set (RS)	1, 5, 10, 15, 20, 25, 50, 100
Query Terms (e)	1, 5, 10, 20, 50, 100, 150, 200, 250, 300, 350, 400, 450
Retrieval Similarity Configurations	$Jelinek\text{-}Mercer$ $(0.1n)$ for $n \in \{0, 1, 2, \dots, 9\}$
	$Dirichlet$ $(500n)$ for $n \in \{1, 2, \dots, 10\}$

Table 4
Assessing modeling capability by comparing non-boosted versus boosted approaches when ranking with depression languages estimated using RMs and BDI-II clinical information. Boosted and non-boosted represent whether terms are boosted with their associated weights or not. For 1, 100, 500, and 1000 writings, the methods have access to these respective quantities of writings per user when ranking. Metrics are defined in 4.2. *RS* stands for relevance set size, *terms* refers to the number of top e selected terms, and *smooth.* denotes the optimal retrieval similarity configuration.

						1 writing			100 writings			500 writings			1000 writings		
	Method	Model	RS	Terms	Smooth.	*P@10*	*NDCG@10*	*NDCG@100*	*P@10*	*NDCG@10*	*NDCG@100*	*P@10*	*NDCG@10*	*NDCG@100*	*P@10*	*NDCG@10*	NDCG@100
Baseline	eRisk 22	Best	–	–	–	0.80	0.88	0.54	0.90	0.93	0.67	0.90	0.92	0.74	0.80	0.86	0.72
		90rh perc.	–	–	–	0.80	0.82	0.44	0.70	0.74	0.59	0.80	0.80	0.61	0.80	0.80	0.62
		Average	–	–	–	0.32	0.34	0.21	0.35	0.36	0.29	0.40	0.41	0.30	0.36	0.39	0.32
	Cross-Encoder	MiniLM-L-6-v2	–	20	–	0.60	0.72	0.34	0.20	0.16	0.35	0.30	0.37	0.36	0.50	0.45	0.39
Our methods	Non-boost.	RM1	80	250	JM(0.9)	0.60	0.66	**0.41**	**1.00**	**1.00**	**0.59**	0.90	0.93	**0.66**	**1.00**	**1.00**	**0.65**
		DMM	20	200	JM(0.9)	**0.70**	0.75	**0.41**	0.90	0.93	0.57	**1.00**	**1.00**	0.60	**1.00**	**1.00**	0.58
		MEDMM	20	200	JM(0.9)	**0.70**	**0.76**	0.40	**1.00**	**1.00**	0.57	**1.00**	**1.00**	0.62	**1.00**	**1.00**	0.59
	Boosted	RM1	80	350	JM(0.9)	0.70	0.76	0.42	**0.90**	**0.91**	**0.59**	**1.00**	**1.00**	**0.63**	**1.00**	**1.00**	**0.61**
		DMM	30	300	JM(0.9)	**0.80**	0.75	**0.43**	0.80	0.86	0.58	0.80	0.86	0.61	0.80	0.86	0.60
		MEDMM	20	200	JM(0.9)	**0.80**	**0.81**	0.42	**0.90**	**0.91**	0.58	0.90	0.94	0.60	**1.00**	**1.00**	0.60

constraint, which limits its ability to process a user's full writing history.

While the transformer-based models offers competitive performance, particularly in the 1-writing scenario, its limitations highlight the trade-offs between accuracy and interpretability in this domain. In contrast, our relevance-based statistical language models provide clinically grounded interpretability: their term weight distributions can be directly aligned with BDI-II symptoms, enabling transparent inspection of the linguistic signals driving the rankings. Moreover, our results show that our probabilistic relevance models generalize more robustly across different quantities of user writings, whereas the transformer baseline suffers from input length constraints and varying performance. We therefore view these probabilistic models not as replacements for neural architectures but as complementary approaches that foreground interpretability, generalization, and transparency while maintaining competitive effectiveness.

These results demonstrate the efficacy of relevance feedback approaches in capturing depression-related language and their potential for early depression detection. Notably, comparing boosted and non-boosted methods reveals that incorporating term weights does not consistently improve performance. This suggests that while boosting can enhance certain retrieval aspects, non-boosted methods are equally, if not more, effective in modeling depression language.

Moreover, to support interpretability, we analyze the learned depression language by identifying terms with the largest differences in weight between the depression and control groups. Rather than report top-k terms by raw frequency, which often favors high-usage generic tokens, we rank terms by their relative weight change between groups. This highlights terms that most clearly differentiate depressive language from typical usage.

Table 5 presents the top 10 most distinctive terms on Reddit. Several of these reflect emotional or social themes. For example, *feel*, *help*, and *friend* suggest emotional disclosure and a search for support, both of which are common in depressive language. The prominence of *im*, *i'll*, and *dont* aligns with prior work showing that depressed individuals often rely more on first-person pronouns and negation, signaling self-focus and negative framing [37,71]. The frequent reliance on absolutist or categorical language (e.g., *dont*, *never*, or related markers) is consistent with Beck's cognitive model of depression [36] and with more recent work identifying absolutist thinking as a distinctive linguistic marker of depressive cognition [38]. Moreover, the term "d" is attributed to the use of past simple in verbs, reflecting the tendency of depressed individuals to reference past events, consistent with theories linking depression to ruminative focus on past experiences [40].

Table 5
Top 10 terms with the largest weight differences between depressed and control users on Reddit. Values reflect percent change relative to control group weights.

Term	% Change
im	801.58%
d	355.74%
i'll	268.91%
game	222.16%
dont	201.15%
friend	198.16%
help	120.08%
she	107.95%
feel	96.14%
also	95.51%

These differential term weights therefore resonate with established psychological and psycholinguistic theories rather than being isolated computational artifacts [37,71]. They help clarify how depressive language manifests on Reddit, while also offering practical insights for researchers and clinicians interested in linguistic signals of mental health. The full depression and control language models are publicly available at https://gitlab.irlab.org/eliseo.bao/rf-models-depression-language.

Table 6
Generalizing with the non-boosted approach for ranking with RM depression languages. The first column represents the platform used to train the RMs. The second column indicates the platform used to test the ranking results. Asterisk symbols denote that hyperparameters are reutilized from the same-platform optimization.

						1 writing			100 writings			500 writings			1000 writings		
Train	Test	Model	RS	Terms	Smooth.	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100
Reddit	Reddit	RM1	80	250	JM(0.9)	0.60	0.66	**0.41**	**1.00**	**1.00**	**0.59**	0.90	0.93	**0.66**	**1.00**	**1.00**	**0.65**
		DMM	20	200	JM(0.9)	**0.70**	0.75	**0.41**	0.90	0.93	0.57	**1.00**	**1.00**	0.60	**1.00**	**1.00**	0.58
		MEDMM	20	200	JM(0.9)	**0.70**	**0.76**	0.40	**1.00**	**1.00**	0.57	**1.00**	**1.00**	0.62	**1.00**	**1.00**	0.59
	Twitter	RM1	*	*	*	0.40	0.49	0.43	**0.70**	**0.76**	**0.59**	**0.90**	**0.94**	**0.64**	**1.00**	**1.00**	**0.62**
		DMM	*	*	*	**0.50**	**0.61**	0.47	0.60	0.71	0.57	0.50	0.62	0.61	0.80	0.86	0.58
		MEDMM	*	*	*	**0.50**	**0.61**	0.48	**0.70**	0.75	0.56	0.50	0.63	0.62	0.80	0.85	0.58
Twitter	Twitter	RM1	3	10	JM(0.1)	0.30	0.20	0.38	**0.90**	**0.89**	**0.58**	0.50	0.62	0.56	0.60	0.69	**0.58**
		DMM	3	5	JM(0.7)	**0.60**	**0.72**	0.41	0.60	0.66	0.54	**0.70**	0.60	**0.57**	**0.90**	**0.78**	0.57
		MEDMM	3	250	JM(0)	0.50	0.57	**0.44**	0.60	0.67	0.53	0.60	**0.71**	0.53	0.50	0.63	0.50
	Reddit	RM1	*	*	*	**0.40**	**0.38**	**0.33**	**0.60**	**0.50**	**0.34**	**0.40**	**0.41**	**0.35**	**0.50**	**0.56**	**0.36**
		DMM	*	*	*	0.10	0.09	0.18	0.30	0.21	0.32	0.30	0.22	0.34	0.40	0.28	0.34
		MEDMM	*	*	*	0.20	0.15	0.26	0.50	0.53	0.32	0.20	0.14	0.31	0.30	0.22	0.34

RQ1 Observation

Non-boosted and boosted approaches deliver similar performance across all writing counts. Notably, multiple models achieved perfect scores of 1.00 for *P*@10 and *P*@20 starting from 100 writings, exceeding the best baseline score for these settings. For other writing counts, the results consistently outperformed the 90th percentile baseline score.

4.4. RQ2: Can the Model be Generalized to Other Social Media Platforms?

Recently, concerns have been raised about the generalization ability of mental health models trained on social media data [20,72], highlighting issues with proxy-based methods for annotating mental health status on these platforms [73]. Specifically, such models often capture dataset-specific features and direct mentions of mental health rather than subtle indications of depression symptoms. This limits their applicability in real-world scenarios and typically leads to performance losses when transferring across platforms.

In the previous subsection, we demonstrated that relevance feedback language models effectively model depression language, producing rankings where users at higher risk are ranked at the top. However, these experiments relied solely on data from Reddit, leaving the generalization ability of our methods unexplored. To address this, we define **RQ2** to evaluate how well our methods perform on out-of-domain data using both non-boosted and boosted approaches. To answer this question, we propose additional experiments using the same setup as in Section 4.3, but incorporating a different platform, Twitter, with the binary depression dataset from CLPsych described in Section 4.1.

4.4.1. Evaluating cross-platform performance

To evaluate cross-platform performance, we used pre-optimized Reddit queries to rank Twitter subjects, maintaining the same evaluation framework as in previous experiments. The results of this setting (i.e., training on Reddit users) for non-boosted and boosted methods are shown in the first blocks of Table 6 and Table 7, respectively. Overall, Reddit queries demonstrate robust performance when generalizing to Twitter, with both non-boosted and boosted methods achieving *NDCG*@100 values above 0.62 for 500 and 1000 writings. These findings suggest strong generalization potential. When comparing these results to testing on Reddit (i.e., both training and testing on the same platform), the non-boosted methods perform better on Reddit (Table 6). Under boosted approaches (Table 7), performance remains consistent across platforms. However, factors such as the positive-to-control user ratio (1 : 2 for Twitter; 1 : 13 for Reddit) complicate direct comparisons. To contextualize Reddit-to-Twitter generalization, we compare these results with baseline rankings on Twitter, where queries are optimized with Twitter data.

As detailed in Section 3, BDI-II scores are crucial for computing relevance feedback language models. However, since the Twitter dataset is binary and lacks BDI-II scores, we assumed a constant score of 1, treating all users equally in depression language estimation. Using this approach, we optimized Twitter queries by determining the best relevance set size, number of query terms, and retrieval similarity hyperparameters, as outlined in Section 4.3. Results for this setup (i.e., training on Twitter users) are reported in the second block of Tables 6 and 7. Under this approach, we achieved *NDCG*@100 values above 0.50 for both non-boosted and boosted methods across 100, 500, and 1000 writings. Surprisingly, Reddit queries outperformed Twitter-optimized queries for Twitter rankings. This counterintuitive result stems from the assumption of a constant BDI-II score for all Twitter users, underscoring the importance of BDI-II scores in weighting users' language contributions. Without this approach, performance degrades significantly.

We also evaluated Twitter queries on Reddit users. In this direction, results were poorer, with *NDCG*@100 scores around 0.30 for both non-boosted and boosted methods. This is again attributed to the assumption of a constant BDI-II score for Twitter users during depression language modeling. These observations emphasize the critical role of validated clinical scores such as BDI-II in training predictive models. This asymmetry in cross-platform performance may also be partly explained by Reddit's longer and more context-rich posts, which provide deeper linguistic signals of depression. In contrast, Twitter's shorter and more fragmented content may limit the ability to model these signals effectively [42], since it can also affect the way people express themselves and their emotions and feelings [74,75].

RQ2 Observation

The evaluation of cross-platform performance shows that pre-optimized Reddit queries effectively rank Twitter subjects, achieving *NDCG*@100 values exceeding 0.62 for 500 and 1000 writings, demonstrating strong generalization potential. However, applying Twitter queries to Reddit results in lower performance, attributed to the lack of BDI-II scores in Twitter data. This highlights the critical role of incorporating clinical markers like BDI-II scores in detection systems.

4.4.2. Integrating data from multiple platforms into a single model

In addition to demonstrating cross-platform transferability, we evaluated the potential of integrating data from multiple platforms into a

Table 7
Generalizing with the boosted approach for ranking with RM depression languages. Asterisk symbols denote that hyperparameters are reutilized from the same-platform optimization.

Train	Test	Model	RS	Terms	Smooth.	1 writing			100 writings			500 writings			1000 writings		
						P@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100
Reddit	Reddit	RM1	80	350	JM(0.9)	0.70	0.76	0.42	**0.90**	**0.91**	**0.59**	**1.00**	**1.00**	**0.63**	**1.00**	**1.00**	**0.61**
		DMM	30	300	JM(0.9)	**0.80**	0.75	**0.43**	0.80	0.86	0.58	0.80	0.86	0.61	0.80	0.86	0.60
		MEDMM	20	200	JM(0.9)	**0.80**	**0.81**	0.42	**0.90**	**0.91**	0.58	0.90	0.94	0.60	**1.00**	**1.00**	0.60
	Twitter	RM1	*	*	*	**0.70**	**0.73**	**0.53**	**0.50**	0.63	**0.59**	**0.90**	**0.93**	**0.65**	**0.90**	**0.93**	**0.62**
		DMM	*	*	*	0.50	0.54	0.50	**0.50**	**0.65**	0.58	**0.90**	0.92	0.64	**0.90**	**0.93**	0.59
		MEDMM	*	*	*	**0.70**	0.72	0.51	**0.50**	0.62	0.56	0.70	0.78	0.62	0.90	0.92	0.59
Twitter	Twitter	RM1	7	10	JM(0)	0.30	0.25	0.38	**0.80**	**0.71**	0.50	0.60	0.47	0.52	0.50	0.59	**0.56**
		DMM	5	450	JM(0)	**0.50**	0.41	0.38	0.50	0.57	**0.51**	0.40	0.50	**0.56**	**0.60**	0.48	0.48
		MEDMM	3	250	JM(0)	**0.50**	**0.42**	**0.40**	0.50	0.56	**0.51**	**0.60**	**0.71**	**0.56**	**0.60**	**0.69**	0.51
	Reddit	RM1	*	*	*	0.10	0.07	**0.30**	**0.60**	**0.56**	**0.32**	**0.60**	**0.50**	**0.32**	**0.50**	**0.43**	**0.31**
		DMM	*	*	*	**0.30**	**0.23**	0.28	0.30	0.26	0.30	0.40	0.34	0.22	0.20	0.29	0.16
		MEDMM	*	*	*	**0.30**	0.22	**0.30**	0.40	0.32	0.31	0.30	0.23	0.21	0.10	0.07	0.14

Table 8
Ranking users from multiple platforms. Best Reddit and Twitter alone results are also reported for clarity and context.

Method	Model	RS	Terms	Smooth.	1 writing			100 writings			500 writings			1000 writings		
					P@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	NDCG@100
Best R-R	RM1	80	250	JM(0.9)	**0.60**	**0.66**	0.41	**1.00**	**1.00**	**0.59**	**0.90**	0.93	**0.66**	**1.00**	**1.00**	**0.65**
Best T-T	RM1	3	10	JM(0.1)	0.30	0.20	0.38	0.90	0.89	0.58	0.50	0.62	0.56	0.60	0.69	0.58
Best R-T	RM1	*	*	*	0.40	0.49	**0.43**	0.70	0.76	**0.59**	**0.90**	**0.94**	0.64	**1.00**	**1.00**	0.62
Best T-R	RM1	*	*	*	0.40	0.38	0.33	0.60	0.50	0.34	0.40	0.41	0.35	0.50	0.56	0.36
Non-boost.	RM1	20	200	JM(0.9)	0.60	0.72	**0.41**	**0.90**	**0.93**	**0.62**	**1.00**	**1.00**	**0.71**	**1.00**	**1.00**	0.68
	DMM	5	150	JM(0.9)	**0.70**	**0.75**	0.40	**0.90**	**0.93**	0.60	**1.00**	**1.00**	0.67	**1.00**	**1.00**	**0.69**
	MEDMM	5	150	JM(0.9)	0.60	0.58	0.38	**0.90**	**0.93**	0.60	**1.00**	**1.00**	0.67	**1.00**	**1.00**	0.68
Boost.	RM1	15	150	JM(0.8)	0.80	**0.80**	0.43	0.80	0.85	**0.61**	**1.00**	**1.00**	0.68	**1.00**	**1.00**	**0.68**
	DMM	3	200	JM(0.9)	0.70	0.78	**0.47**	0.90	0.92	0.53	0.90	0.93	0.61	0.90	0.93	0.60
	MEDMM	15	250	JM(0.9)	**0.80**	0.77	0.42	**1.00**	**1.00**	0.60	**1.00**	**1.00**	**0.69**	**1.00**	**1.00**	0.66

single model. By incorporating linguistic features from diverse domains, we aimed to enhance the model's ability to capture varied language patterns, resulting in improved overall performance. This approach is designed to create models capable of effectively ranking users regardless of the platform on which they write. To this end, we conducted an experiment combining Reddit and Twitter data. Specifically, we used the Reddit BDI-II dataset and the Twitter binary depression dataset to estimate new depression languages. Since the Twitter dataset lacks BDI-II scores, we assumed a constant score of 46 for all Twitter users, corresponding to the median of the severe depression category. Using these combined datasets, we estimated new depression languages and repeated the ranking and evaluation frameworks. Results for both non-boosted and boosted approaches are shown in Table 8. Performance metrics indicate that this multi-domain integration outperforms the best single-platform methods across all writing counts. From 100 writings onward, *P*@10 and *NDCG*@100 values are perfect or near-perfect. *NDCG*@100 reaches its highest value of 0.71 for 500 writings using the non-boosted approach, a 0.05-point improvement over the best single-platform result.

We also revisited the assumption of a constant BDI-II score for Twitter users. Previously, when using only Twitter data, this assumption led to poor performance. However, when integrating data from multiple platforms, even with a uniform BDI-II score for Twitter users, the inclusion of Twitter data broadens linguistic diversity and enhances results. The models benefit from diverse input data, with Twitter contributing to language variation and Reddit providing weighted term estimates based on BDI-II scores. These results highlight the potential of integrating knowledge from multiple platforms into a single model, particularly for scenarios where some platforms lack training data or clinical annotations like BDI-II scores. This approach offers a robust framework for ranking users across various platforms while leveraging the strengths of each data source.

RQ2 Observation

Integrating data from multiple platforms into a single model enables the capture of diverse linguistic features, leading to improved overall performance. This approach is especially valuable when high-quality training data, such as clinical markers, is unavailable for certain platforms.

4.5. RQ3: *Can Relevance-Based Statistical Language Models Enhance Existing Lexicons?*

For **RQ1** and **RQ2**, we formulated ranking methods based on queries derived from the top-weighted terms identified by our RMs to capture depression language. However, the depression languages and term selection explored thus far have not been supervised by domain experts. To address this, we investigate whether incorporating knowledge from established clinical depression lexicons can enhance the performance of our ranking methods. Specifically, we incorporate the Pedesis and De Choudhury depression lexicons, which are collections of terms identified as indicators of depression in social media texts [8,29]. These resources are widely used in computational studies to analyze language patterns and detect depression using NLP techniques.

The core of this experiment involves assigning our RMs depression language weights to the terms in the lexicons. With these weights, we

Table 9
Baseline rankings. Queries defined in 4.5.1.

Query	Smooth	1 writing			100 writings			500 writings			1000 writings		
		P@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100	*P*@10	*NDCG*@10	*NDCG*@100
Best RM query		0.60	0.66	0.41	1.00	1.00	0.59	0.90	0.93	0.66	1.00	1.00	0.65
Q1	D(1500)	0.10	0.22	0.07	**0.80**	**0.82**	0.39	0.60	0.67	0.41	0.60	0.69	0.40
Q2	D(500)	0.30	0.36	**0.30**	0.70	0.70	**0.57**	0.80	0.85	**0.58**	**0.80**	**0.86**	**0.60**
Q3	JM(0)	0.30	0.25	0.27	0.20	0.16	0.26	0.30	0.20	0.20	0.20	0.16	0.16
Q4	D(500)	**0.80**	**0.78**	**0.30**	0.70	0.76	0.44	**0.90**	0.93	0.46	**0.80**	0.85	0.47
Q5	D(500)	0.40	0.35	**0.30**	0.70	0.80	0.48	0.80	0.71	0.43	0.40	0.40	0.38
Q6	JM(0.9)	0.00	0.00	0.20	0.10	0.08	0.19	0.20	0.23	0.22	0.20	0.20	0.22
Q7	D(500)	0.40	0.42	0.24	0.50	0.61	0.42	**0.90**	**0.94**	0.44	0.70	0.79	0.43
Q8	JM(0.9)	0.30	0.25	0.25	0.40	0.46	0.35	0.70	0.65	0.33	0.50	0.47	0.28
Q9	JM(0.2)	0.30	0.25	0.27	0.30	0.26	0.31	0.30	0.30	0.29	0.30	0.30	0.24

Table 10
Ranking performance of the Pedesis and De Choudhury clinical lexicons when incorporating the weights estimated by our relevance feedback models as part of the strategy for selecting terms when deriving queries. This experiment is run under the evaluation framework introduced in § RQ1. Results for the best-performing query derived from the depression lexicons estimated with relevance feedback models and the Pedesis and De Choudhury clinical lexicons without information about the weight of each term (Q2 and Q3) are reported as reference baselines.

Best RM query		RM1	80	250	JM(0.9)	0.60	0.66	0.41	1.00	1.00	0.59	0.90	0.93	0.66	1.00	1.00	0.65
	Q2 Baseline	–	–	–	D(500)	0.30	0.36	0.30	0.70	0.70	0.57	0.80	0.85	0.58	0.80	0.86	0.60
Choudhury	Non-boost.	RM1	15	10	JM(0.6)	**0.50**	**0.65**	**0.34**	**1.00**	**1.00**	**0.62**	**0.90**	**0.94**	**0.65**	**0.90**	**0.94**	**0.65**
		DMM	5	100	D(100)	0.30	0.24	0.28	0.60	0.64	0.56	0.40	0.52	0.57	0.60	0.65	0.57
		MEDMM	20	20	JM(0.8)	0.30	0.30	0.23	0.80	0.84	0.56	0.80	0.86	0.64	0.80	0.86	0.62
	Boost.	RM1	20	10	JM(0.6)	0.20	0.20	0.21	0.80	0.85	0.58	**0.80**	**0.84**	0.63	**0.80**	**0.84**	**0.64**
		DMM	20	100	D(500)	**0.60**	**0.48**	0.28	0.70	0.70	**0.60**	0.70	0.72	**0.64**	0.60	0.62	0.60
		MEDMM	20	50	D(500)	0.50	0.45	**0.29**	**0.90**	**0.90**	0.59	**0.80**	0.79	0.61	**0.80**	0.77	0.61
	Q3 Baseline	–	–	–	JM(0)	0.30	0.25	0.27	0.20	0.16	0.26	0.30	0.20	0.20	0.20	0.16	0.16
Pedesis	Non-boost.	RM1	30	5	JM(0.3)	**0.40**	0.44	0.30	**0.60**	**0.65**	**0.49**	**0.60**	**0.65**	**0.55**	**0.70**	**0.73**	**0.52**
		DMM	15	5	D(500)	**0.40**	**0.54**	**0.36**	0.40	0.44	0.46	0.50	0.53	0.47	0.40	0.46	0.48
		MEDMM	3	5	D(500)	**0.40**	**0.54**	**0.36**	0.40	0.44	0.46	0.50	0.53	0.47	0.40	0.46	0.48
	Boost.	RM1	30	5	JM(0.1)	**0.60**	**0.48**	**0.30**	**0.70**	**0.69**	**0.53**	**0.60**	**0.65**	**0.58**	**0.60**	**0.65**	**0.56**
		DMM	30	5	JM(0.2)	0.50	0.42	0.29	0.60	0.67	0.48	0.50	0.61	0.56	0.50	0.61	0.55
		MEDMM	30	5	JM(0.3)	0.50	0.43	**0.30**	0.60	0.68	0.49	0.50	0.63	0.55	0.50	0.63	0.55

determine the relative importance of each lexicon term in the context of depression, enabling us to design strategies for selecting top terms for ranking queries. To evaluate the effectiveness of this approach, we first assess the lexicons' baseline performance without incorporating our weights, as detailed in Section 4.5.1. This allows us to later measure the impact of incorporating RMs depression language weights on ranking performance.

4.5.1. Baseline queries

We defined a set of 9 queries derived from the Pedesis and De Choudhury lexicons. These queries were constructed using unique terms and adjectives, expanded using the methods proposed by Losada and Gamallo [76]:

Q1 : {*sad*, *lonely*, *hopeless*, *worthless*}[8]
Q2 : {All unique terms (106) from De Choudhury}
Q3 : {All unique terms (636) from Pedesis}
Q4 : {Unambiguous adjectives (7) from De Choudhury}
Q5 : {Unambiguous adjectives (153) from Pedesis}
Q6 : {*DE*[9] expanded unamb. adj. (13) from De Choudhury}
Q7 : {*WE* expanded unamb. adj. (16) from De Choudhury}
Q8 : {*DE* expanded unamb. adj. (312) from Pedesis}
Q9 : {*WE* expanded unamb. adj. (549) from Pedesis}

Using these queries and the evaluation framework from previous experiments, we generated an initial baseline ranking for the Reddit binary depression dataset from eRisk 2022. Results are presented in Table 9, alongside scores from the best-performing query derived from RM depression languages. For *P*@10 and *NDCG*@10, queries **Q4** and **Q7** show improvements for 1 and 500 writings compared to the best RM-derived terms. However, for *NDCG*@100, all nine baseline queries produced lower scores than the best-performing RM query.

4.5.2. Ranking with lexicons and BDI-II RF model weights

After computing the baseline queries, we integrated clinically validated depression-related lexicons into our ranking pipeline. The Pedesis lexicon and De Choudhury lexicon include relevant terms validated by clinical experts in the context of depression. However, these lexicons lack weights or relevance scores for their terms. In contrast, our relevance feedback models estimate term weights for depression languages. In this experiment, we assign these estimated weights to the terms in the clinical lexicons to derive new depression queries and generate rankings, following the experimental and evaluation frameworks from Section 4.3. As in previous experiments, we explore two strategies: using weights as query term boosts and using weights to select terms for queries.

Results are presented in Table 10, alongside: (i) the best-performing query derived from our relevance feedback models (**RQ1**) and (ii) the baseline results from the Pedesis and De Choudhury clinical lexicons without term weights (**Q2**/**Q3**).

[8] The query is not defined *ad hoc*, but rather taken from the work of Losada and Gamallo [76]. These are the four words the authors identified for the vector associated with depression.

[9] *DE* (Distributional-based) and *WE* (WordNet-based) expansions were proposed by Losada and Gamallo [76].

The results show a significant performance improvement when incorporating weights from our relevance feedback models into the query derivation process. **Q2** and **Q3** baseline values are improved across all metrics. Compared to the best-performing query derived solely from relevance feedback models, the new lexicon-based queries achieve comparable results. For example, the non-boosted Choudhury lexicon achieves an *NDCG*@100 value of 0.65 for 1000 writings, matching the best-performing relevance feedback model query. In some cases, the lexicon-based queries surpass relevance feedback-only results, such as the *NDCG*@100 value of 0.65 achieved with the non-boosted Choudhury lexicon for 100 writings. Comparing the two clinical lexicons, results indicate that queries derived from De Choudhury generally outperform those from Pedesis across metrics and settings. For instance, the Choudhury lexicon reaches an *NDCG*@100 value of 0.65 for 1000 writings, while the Pedesis lexicon's highest *NDCG*@100 value is 0.58, obtained with the boosted RM1 method for 500 writings. These findings highlight the superior effectiveness of the Choudhury lexicon in generating depression-related queries when combined with relevance feedback model weights.

RQ3 Observation

Integrating relevance feedback model term weights into clinical lexicons significantly enhances query performance across all metrics. While domain-expert knowledge in language modeling is typically compiled into unsorted lexicons, our experiments demonstrate that incorporating relevance feedback model term weights into the term selection strategy for query derivation has a positive impact on performance.

4.6. RQ4: What Insights Do These Vocabularies Provide?

Our previous research questions were exploratory, focusing on ranking experiments evaluated through offline metrics. In contrast, **RQ4** is an analytical case study aimed at examining our relevance feedback models to gain insights into how individuals with depression use language compared to a control group. The relationship between language and clinical disorders has been extensively studied in text and social analytics [8,77]. One prominent tool in this field is the Linguistic Inquiry and Word Count (LIWC) 2007 [78], which provides a dictionary of approximately 4500 words categorized into 64 psychological categories, including *sadness*, *anxiety*, and *family*. Another well-known resource is the National Research Council of Canada (NRC) Emotion Lexicon [79], which classifies around 14000 words into 8 core emotions (*joy*, *sadness*, *anticipation*, *anger*, *fear*, *disgust*, *trust*, and *surprise*) and two sentiment categories, *positive* and *negative*. Words in the NRC lexicon can belong to multiple categories based on their emotional connotations. While LIWC covers a broad spectrum of psychological and linguistic categories, NRC is particularly effective at quantifying emotional content in language.

In this research question, we leverage LIWC and NRC alongside our relevance feedback models to analyze and visualize language differences between depressed and control users across two platforms, Reddit and Twitter. Our analysis focuses on two objectives: (i) Investigating differences in language use between depressed and control groups, with an emphasis on psychological and emotional categories identified by LIWC and NRC. (ii) Examining platform-specific language trends to identify distinctions between Reddit and Twitter.

Our relevance-based statistical language models assign weights to each term in the generated vocabularies, allowing us to evaluate term importance across categories for both control and depressed groups within the external lexicons of LIWC and NRC. To do this, we calculated the cumulative term weights within each lexicon category, with higher cumulative weights indicating greater importance within that group. Fig. 2 illustrates these differences in six LIWC categories (Fig. 2(a)) and the eight core emotions from NRC (Fig. 2(b)). For these categories, Fig. 2 displays cumulative term weights across four groups: Reddit depression, Reddit control, Twitter depression, and Twitter control.

LIWC includes 64 categories, and for this analysis, we selected a sample encompassing both linguistic elements (e.g., verbs and prepositions) and emotionally charged topics (e.g., *friends*, *anxiety*, *religion*, and *death*), as shown in Fig. 2(a). Linguistic categories such as verbs and prepositions show minimal differences across groups, regardless of condition (depressed or control) or platform (Twitter or Reddit). This aligns with our hypothesis that basic language elements are less likely to reveal distinctions between groups. In contrast, emotional categories (*Friends*, *Anxiety*, *Religion*, and *Death*) exhibit notable differences between depressed and control groups, with emotional terms carrying consistently higher weights for depressed users. Moreover, terms related to these emotional categories are particularly prominent in depressed users on Reddit.

Fig. 3 and Fig. 4 offer a detailed view of the data using violin plots, which visualize the distribution and density of term weights within each category [43], as well as the median and range. The top plot (Fig. 3(a)) corresponds to the Reddit dataset, while the bottom plot (Fig. 3(b)) presents the analysis for the Twitter dataset. Across both platforms, the distributions for *verbs* and *prepositions* are similar between depressed and control groups, reinforcing the observation that these general linguistic categories do not reveal significant differences.

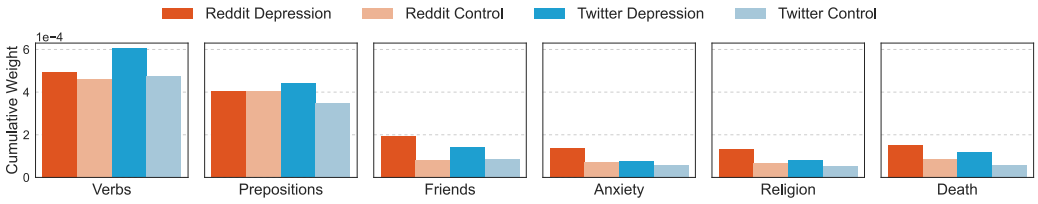
In contrast, the emotional categories (*friends*, *anxiety*, *religion*, and *death*) show distinct patterns when comparing Reddit and Twitter collections. On Reddit, the language of depressed users exhibits a denser concentration of terms within these categories, highlighting their importance among this group. These elevated term weights are consistent with the trends in Fig. 2(a), emphasizing the role of emotional language for depressed Reddit users. By comparison, the Twitter dataset displays less pronounced differences between depressed and control groups. This stronger differentiation on Reddit can be attributed to two primary factors: First, the nature of the data on each platform plays a crucial role. On Twitter, user classification is binary (depressed vs. control), meaning all depressed users contribute equally to the model, regardless of the severity of their condition. Conversely, the Reddit dataset includes BDI-II scores, enabling relevance feedback models to weight users' language contributions based on their depression severity. This results in Reddit models capturing more nuanced patterns associated with varying levels of depression, while the binary classification on Twitter limits the ability to distinguish between groups. Second, Twitter data tends to be noisier due to the prevalence of retweets, which often include external content from public figures or viral tweets that may not reflect the sharer's personal emotional state. Additionally, Twitter's character limit (formerly 140 characters) restricts the depth of expression, whereas Reddit posts are not length-constrained. These factors dilute the emotional signal from depressed Twitter users, making it harder to differentiate their language from that of control users.

RQ4 Observation

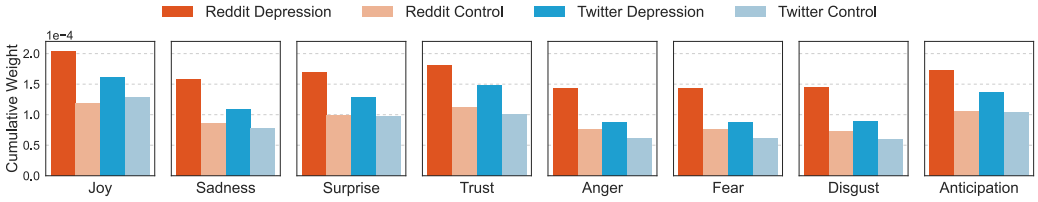
The analysis of language use between depressed and control groups reveals significant differences, especially in emotionally charged categories. The LIWC and NRC lexicons show that depressed individuals have higher term weights in categories like *friends*, *anxiety*, *religion*, and *death*. This pattern is more pronounced on Reddit, likely due to the availability of nuanced depression severity data compared to the binary classification and noisier data from Twitter.

5. Conclusions

This work aimed to: (i) assess the effectiveness of relevance-based statistical language models for modeling depression language, (ii) evaluate the generalizability of the method across different social media

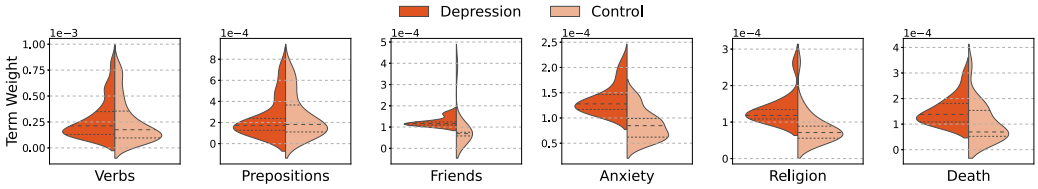


(a) LIWC lexicon.

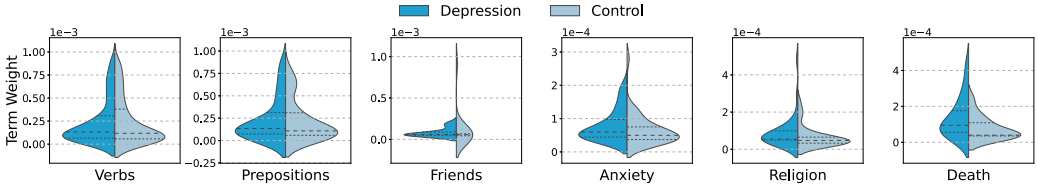


(b) NRC lexicon.

Fig. 2. Cumulative term weights across various lexicon categories for Reddit and Twitter vocabularies. The color scheme distinguishes between the vocabularies of depressed and control groups on each platform.

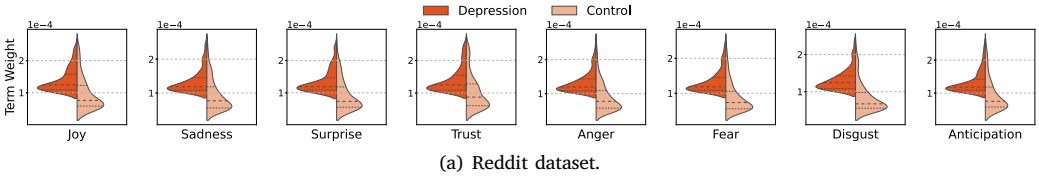


(a) Reddit dataset.

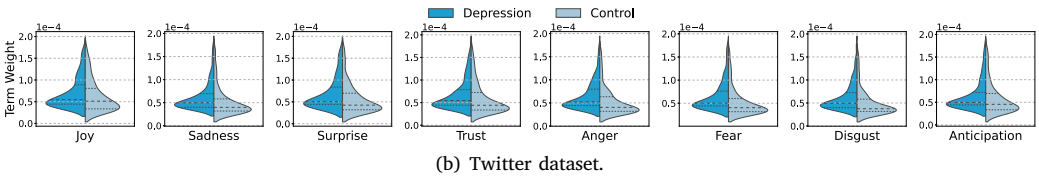


(b) Twitter dataset.

Fig. 3. Term weight distribution across various LIWC categories for Reddit (a) and Twitter (b) datasets. The color scheme differentiates between the vocabularies of depressed and control groups.



(a) Reddit dataset.



(b) Twitter dataset.

Fig. 4. Term weight distribution across various NRC categories for Reddit (a) and Twitter (b) datasets. The color scheme differentiates between the vocabularies of depressed and control groups.

platforms, (iii) explore potential improvements by incorporating established depressive lexicons, and (iv) analyze depression language to gain insights into the vocabulary used by individuals with depression disorders. The experimental results allow us to draw the following conclusions:

Relevance-based statistical language models effectively capture the

nuances of depression language. Both boosted and non-boosted ranking approaches demonstrated their ability to rank users by estimated depression risk based on social media writings. These models showed strong potential for early detection, achieving strong results with only 100 writings per user. The method's generalizability was demonstrated through its application to Twitter data alongside the original Reddit dataset. Models trained on Reddit data maintained robust performance when applied to Twitter users, achieving high ranking scores. Moreover, integrating data from multiple platforms into a single model improved performance. By combining Reddit and Twitter data, the models leveraged diverse linguistic features, enhancing their ability to detect depression signs, particularly on platforms lacking clinical training data with scores.

Incorporating domain-expert validated knowledge from the Pedesis and De Choudhury lexicons further improved the system's effectiveness. Combining expert-validated terms with relevance-based weighting significantly enhanced the ranking task, demonstrating the value of integrating established lexicons. Finally, the analysis of language use between depressed and control groups using LIWC and NRC lexicons revealed that terms associated with emotional content (e.g., *friends*, *anxiety*, *religion*, *death*) carried higher weights for depressed users. This difference was more pronounced in Reddit data, likely due to the nuanced severity information provided by BDI-II scores, compared to the binary classification and noisier data from Twitter. Although transformer-based models demonstrate strong predictive capabilities, our results show that relevance-based statistical language models offer a complementary value by foregrounding interpretability, clinical alignment, and cross-platform robustness, making them particularly well-suited for integration into mental health screening pipelines. The distinctive linguistic patterns revealed by our models align with established psychological theories of depression. For instance, increased self-referential language reflects self-focused attention, while negation and absolutist terms are consistent with cognitive models of depressive thinking, reinforcing that our findings capture substantive psychological signals rather than computational artifacts.

While our method is not intended as a standalone diagnostic system, it could be integrated as part of a broader pipeline for digital mental health screening. For example, on social media platforms, it could help identify language patterns that exhibit signals commonly associated with depression. These alerts would not replace professional judgment but could serve as a triage mechanism, allowing clinicians or platform safety teams to prioritize cases that may require timely attention.

Recent regulatory shifts have increased the urgency of developing such tools. In the European Union, the Digital Services Act (DSA) requires very large online platforms to assess and mitigate risks to users' mental well-being [80]. Similar legislative efforts, such as the Kids Online Safety Act (KOSA) in the United States, aim to protect vulnerable groups like minors from exposure to suicide-related or harmful content [81]. In response, major platforms have begun deploying early-stage interventions. Meta uses AI systems alongside user reports to detect potential self-harm content and deliver support resources [82]. Reddit's "Reddit Cares" allows users to report concern for others' well-being, triggering a private message with crisis support links [83]. X (formerly Twitter) provides notifications and contact information for support organizations in response to relevant search terms [84].

Our model could contribute to these kinds of systems by improving early detection and triage accuracy in a privacy-aware and scalable way. Future work should explore deployment challenges such as algorithmic fairness, user consent, and data governance. Ensuring that models do not disproportionately misclassify or overlook vulnerable groups is essential for equitable use, while respecting user autonomy and clarifying the conditions of data use are critical for trust. These concerns are especially important in the context of public accountability and compliance with regulations like the GDPR, the EU Digital Services Act, and the U.S. Kids Online Safety Act. Ultimately, we envision relevance-based language models as one layer in ethically designed, human-in-the-loop systems that promote early intervention and reduce barriers to care.

CRediT authorship contribution statement

Eliseo Bao: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Investigation, Data curation. **Anxo Perez:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Formal analysis, Conceptualization. **David Otero:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Javier Parapar:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Limitations

This study has limitations that should be acknowledged. First, our approach is restricted to textual features, which aligns with our focus on modeling depression language via clinically grounded instruments such as the BDI-II [33]. While this design enables interpretable modeling of linguistic signals, it necessarily omits interactional and relational features (e.g., replies, likes, social reciprocity, or network structures) that are known to influence how depression manifests in online settings. The datasets used in this work do not include such features, and the eRisk evaluation framework defines baselines solely on textual content, which constrains the scope of our comparisons. Nevertheless, our approach is scalable and complementary: the proposed statistical language models could be readily integrated with future systems that incorporate relational or behavioral features, thereby capturing a fuller picture of depression-related online behaviors.

Second, our experiments are conducted on English Reddit and Twitter data, which may limit generalizability to other languages, platforms, or cultural contexts. While our cross-platform results show robustness, extending the approach to multilingual or multimodal settings remains an important direction for future research. Finally, while our models provide clinically aligned interpretability, they are not designed as diagnostic tools. As commented in our conclusions, their outputs should be used as part of broader, human-in-the-loop systems where professional judgment remains central. Moreover, our engagement with ethical challenges such as fairness, consent, and governance remains preliminary, and future work must explore these dimensions more thoroughly to ensure responsible and equitable deployment in real-world contexts.

Ethical statement

The data used in this research were obtained from publicly accessible repositories, in compliance with the exempt status outlined in Title 45 CFR §46.104. All datasets were used in strict adherence to their respective data usage policies. To ensure privacy, robust measures were implemented to maintain anonymity and prevent the identification of personal information. The datasets, sourced from Reddit, were utilized in full accordance with the platform's terms of use. It is important to emphasize that the systems described in this study are intended to support healthcare professionals, not replace them. The development of such technologies requires a cautious approach, prioritizing ethical deployment and maintaining a strong commitment to user privacy, autonomy, and fairness. Addressing consent, algorithmic bias, and data governance will be essential for ensuring that future applications of this work are aligned with societal expectations and regulatory frameworks.

Funding

This work has received support from projects: PLEC2021-007662 (MCIN/AEI/10.13039/501100011033 Ministerio de Ciencia e Innovación, European Union NextGeneration) and PID2022-137061OB-C21 (MCIN/AEI/10.13039/501100011033/, Ministerio de Ciencia e Innovación, by the European Union); Consellería de Educación, Universidade e Formación Profesional, Spain (grant number ED481A-2024-079

and accreditation 2019–2022 ED431G/01 and GRC ED431C 2025/49) and the European Regional Development Fund, which acknowledges the CITIC Research Center.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Eliseo Bao, Anxo Perez, David Otero, Javier Parapar reports financial support was provided by Spain Ministry of Science and Innovation. Eliseo Bao reports financial support was provided by Government of Galicia Department of Education Science Universities and Professional Training. The authors wish to declare the following affiliations and relationships that could be considered as conflicts of interest: - University of A Coruña (udc.es) - University of Santiago de Compostela (usc.es) - Technical University of Darmstadt (tu-darmstadt.de) If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] M. Prince, V. Patel, S. Saxena, M. Maj, J. Maselko, M.R. Phillips, A. Rahman, No health without mental health, Lancet 370 (9590) (2007) 859–877, http://dx.doi.org/10.1016/S0140-6736(07)61238-0.

[2] Australian Institute of Health and Welfare, Mental Health: Prevalence and Impact, Tech. Rep., Australian Institute of Health and Welfare, 2022.

[3] World Health Organization, et al., Depression and Other Common Mental Disorders: Global Health Estimates, Tech. Rep., World Health Organization, 2017.

[4] A. Picardi, I. Lega, L. Tarsitani, M. Caredda, G. Matteucci, M. Zerella, R. Miglio, A. Gigantesco, M. Cerbo, A. Gaddini, F. Spandonaro, M. Biondi, The SET-DEP Group, A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care, J. Affect. Disord. 198 (2016) 96–101, http://dx.doi.org/10.1016/j.jad.2016.03.025.

[5] A. Gulliver, K.M. Griffiths, H. Christensen, Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review, BMC Psychiatry 10 (1) (2010) 113, http://dx.doi.org/10.1186/1471-244X-10-113.

[6] G. Thornicroft, Stigma and discrimination limit access to mental health care, Epidemiologia E Psichiatr. Soc. 17 (1) (2008) 14–19, http://dx.doi.org/10.1017/S1121189X00002621.

[7] M. Dixon-Woods, D. Cavers, S. Agarwal, E. Annandale, A. Arthur, J. Harvey, R. Hsu, S. Katbamna, R. Olsen, L. Smith, R. Riley, A.J. Sutton, Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups, BMC Med. Res. Methodol. 6 (1) (2006) 35, http://dx.doi.org/10.1186/1471-2288-6-35.

[8] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, Proc. Int. AAAI Conf. Web Social Media 7 (1) (2021) 128–137, http://dx.doi.org/10.1609/icwsm.v7i1.14432.

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

[10] A. Callahan, K. Inckle, Cybertherapy or psychobabble? A mixed methods study of online emotional support, Br. J. Guid. Couns. 40 (3) (2012) 261–278, http://dx.doi.org/10.1080/03069885.2012.681768.

[11] G. Ferraro, B. Loo Gee, S. Ji, L. Salvador-Carulla, Lightme: analysing language in internet support groups for mental health, Heal. Inf. Sci. Syst. 8 (1) (2020) 34, http://dx.doi.org/10.1007/s13755-020-00115-7.

[12] A.-S. Uban, B. Chulvi, P. Rosso, An emotion and cognitive based analysis of mental health disorders from social media data, Future Gener. Comput. Syst. 124 (2021) 480–494, http://dx.doi.org/10.1016/j.future.2021.05.032, URL https://www.sciencedirect.com/science/article/pii/S0167739X21001825.

[13] A.G. Reece, A.J. Reagan, K.L.M. Lix, P.S. Dodds, C.M. Danforth, E.J. Langer, Forecasting the onset and course of mental illness with Twitter data, Sci. Rep. 7 (1) (2017) 13006, http://dx.doi.org/10.1038/s41598-017-12961-9.

[14] J. Parapar, P. Martín-Rodilla, D.E. Losada, F. Crestani, eRisk 2023: Depression, pathological gambling, and eating disorder challenges, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 585–592, http://dx.doi.org/10.1007/978-3-031-28241-6_67.

[15] A. Zirikly, D. Atzil-Slonim, M. Liakata, S. Bedrick, B. Desmet, M. Ireland, A. Lee, S. MacAvaney, M. Purver, R. Resnik, A. Yates (Eds.), Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics, Seattle, USA, 2022.

[16] C.G. Walsh, B. Chaudhry, P. Dua, K.W. Goodman, B. Kaplan, R. Kavuluru, A. Solomonides, V. Subbian, Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence, JAMIA Open 3 (1) (2020) 9–15, http://dx.doi.org/10.1093/jamiaopen/ooz054.

[17] T.U. Hauser, V. Skvortsova, M. De Choudhury, N. Koutsouleris, The promise of a model-based psychiatry: building computational models of mental ill health, Lancet Digit. Heal. 4 (11) (2022) e816–e828, http://dx.doi.org/10.1016/s2589-7500(22)00152-2.

[18] A.T. Beck, R.A. Steer, G. Brown, Beck depression inventory–II, in: PsycTESTS Dataset, American Psychological Association (APA), 1996, http://dx.doi.org/10.1037/t00742-000.

[19] K. Kroenke, R.L. Spitzer, J.B. Williams, The PHQ-9: validity of a brief depression severity measure, J. Gen. Intern. Med. 16 (9) (2001) 606–613.

[20] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, A. Cohan, Improving the generalizability of depression detection by leveraging clinical questionnaires, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8446–8459, http://dx.doi.org/10.18653/v1/2022.acl-long.578.

[21] F. Cacheda, D. Fernandez, F.J. Novoa, V. Carneiro, Early detection of depression: Social network analysis and random forest techniques, J. Med. Internet Res. 21 (6) (2019) e12554, http://dx.doi.org/10.2196/12554.

[22] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 495–503, http://dx.doi.org/10.1145/3159652.3159725.

[23] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, S. Ananiadou, MentaLLaMA: Interpretable mental health analysis on social media with large language models, in: Proceedings of the ACM on Web Conference 2024, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 4489–4500, http://dx.doi.org/10.1145/3589334.3648137.

[24] Y. Hua, F. Liu, K. Yang, Z. Li, Y. han Sheu, P. Zhou, L.V. Moran, S. Ananiadou, A. Beam, Large language models in mental health care: a scoping review, 2024, arXiv:2401.02984.

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

[26] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7184–7190.

[27] N.V. Babu, E.G.M. Kanaga, Sentiment analysis in social media data for depression detection using artificial intelligence: A review, SN Comput. Sci. 3 (1) (2021) 74, http://dx.doi.org/10.1007/s42979-021-00958-1.

[28] J. Novikova, K. Shkaruta, DECK: Behavioral tests to improve interpretability and generalizability of BERT models detecting depression from text, 2022, arXiv:2209.05286.

[29] Y. Neuman, Y. Cohen, D. Assaf, G. Kedma, Proactive screening for depression through metaphorical and automatic text analysis, Artif. Intell. Med. 56 (1) (2012) 19–25, http://dx.doi.org/10.1016/j.artmed.2012.06.001.

[30] V. Adarsh, P. Arun Kumar, V. Lavanya, G. Gangadharan, Fair and explainable depression detection in social media, Inf. Process. Manage. 60 (1) (2023) 103168, http://dx.doi.org/10.1016/j.ipm.2022.103168.

[31] L. Ansari, S. Ji, Q. Chen, E. Cambria, Ensemble hybrid learning methods for automated depression detection, IEEE Trans. Comput. Soc. Syst. 10 (1) (2023) 211–219, http://dx.doi.org/10.1109/TCSS.2022.3154442.

[32] Y.J. Msosa, A. Grauslys, Y. Zhou, T. Wang, I. Buchan, P. Langan, S. Foster, M. Walker, M. Pearson, A. Folarin, A. Roberts, S. Maskell, R. Dobson, C. Kullu, D. Kehoe, Trustworthy data and AI environments for clinical prediction: Application to crisis-risk in people with depression, IEEE J. Biomed. Heal. Informatics 27 (11) (2023) 5588–5598, http://dx.doi.org/10.1109/JBHI.2023.3312011.

[33] A.T. Beck, C.H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Arch. Gen. Psychiatry 4 (6) (1961) 561–571, http://dx.doi.org/10.1001/archpsyc.1961.01710120031004.

[34] L.S. Radloff, The CES-D scale: A self-report depression scale for research in the general population, Appl. Psychol. Meas. 1 (3) (1977) 385–401, http://dx.doi.org/10.1177/014662167700100306.

[35] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, American Psychiatric Association, 2013, http://dx.doi.org/10.1176/appi.books.9780890425596.

[36] A. Beck, Cognitive Therapy of Depression, in: The Guilford Clinical Psychology and Psychotherapy Series, Guilford Press, 1979.

[37] T. Pyszczynski, K. Holt, J. Greenberg, Depression, self-focused attention, and expectancies for positive and negative future life events for self and others, J. Pers. Soc. Psychol. 52 (5) (1987) 994–1001, http://dx.doi.org/10.1037/0022-3514.52.5.994.

[38] M. Al-Mosaiwi, T. Johnstone, In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation, Clin. Psychol. Sci. 6 (4) (2018) 529–542, http://dx.doi.org/10.1177/2167702617747074.

[39] D.E. Losada, F. Crestani, A test collection for research on depression and language use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2016, pp. 28–39.

[40] N. Ramirez-Esparza, C. Chung, E. Kacewic, J. Pennebaker, The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches, Proc. Int. AAAI Conf. Web Social Media 2 (1) (2021) 102–108, http://dx.doi.org/10.1609/icwsm.v2i1.18623.

[41] E.-M.G. Stephanie Rude, J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cogn. Emot. 18 (8) (2004) 1121–1133, http://dx.doi.org/10.1080/02699930441000030.

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

[43] E.A. Ríssola, M. Aliannejadi, F. Crestani, Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation, Inf. Process. Manage. 59 (3) (2022) 102890, http://dx.doi.org/10.1016/j.ipm.2022.102890.

[44] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal To Clinical Reality, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 1–10, http://dx.doi.org/10.3115/v1/W15-1201.

[45] A. Bucur, A. Cosma, L.P. Dinu, Early risk detection of pathological gambling, self-harm and depression using BERT, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - To - 24th, 2021, in: CEUR Workshop Proceedings, vol. 2936, CEUR-WS.org, 2021, pp. 938–949.

[46] A. Murarka, B. Radhakrishnan, S. Ravichandran, Classification of mental illnesses on social media using RoBERTa, in: E. Holderness, A. Jimeno Yepes, A. Lavelli, A.-L. Minard, J. Pustejovsky, F. Rinaldi (Eds.), Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, 2021, pp. 59–68, online.

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

[48] A.U. Kauer, V.P. Moreira, Using information retrieval for sentiment polarity prediction, Expert Syst. Appl. 61 (2016) 282–289, http://dx.doi.org/10.1016/j.eswa.2016.05.038.

[49] A. Htait, S. Fournier, P. Bellot, L. Azzopardi, G. Pasi, Using sentiment analysis for pseudo-relevance feedback in social book search, in: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, ICTIR '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 29–32, http://dx.doi.org/10.1145/3409256.3409847.

[50] F. Najar, N. Bouguila, On smoothing and scaling language model for sentiment based information retrieval, Adv. Data Anal. Classif. 17 (3) (2023) 725–744, http://dx.doi.org/10.1007/s11634-022-00522-6.

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

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

[53] A. Bucur, A. Cosma, L.P. Dinu, P. Rosso, An end-to-end set transformer for user-level classification of depression and gambling disorder, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - To - 8th, 2022, in: CEUR Workshop Proceedings, vol. 3180, CEUR-WS.org, 2022, pp. 851–863.

[54] J.W. Pennebaker, M.R. Mehl, K.G. Niederhoffer, Psychological aspects of natural language. use: our words, our selves, Annu. Rev. Psychol. 54 (2002) 547–577.

[55] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, USA, 2008.

[56] C. Carpineto, G. Romano, A survey of automatic query expansion in information retrieval, ACM Comput. Surv. 44 (1) (2012) http://dx.doi.org/10.1145/2071389.2071390.

[57] J. Rocchio, Relevance feedback information retrieval, Smart Retr. System-Experiments Autom. Document Process. (1971) 313–323.

[58] I. Ruthven, M. Lalmas, A survey on the use of relevance feedback for information access systems, Knowl. Eng. Rev. 18 (2) (2003) 95–145, http://dx.doi.org/10.1017/S0269888903000638.

[59] J.M. Ponte, W.B. Croft, A language modeling approach to information retrieval, in: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98, Association for Computing Machinery, New York, NY, USA, 1998, pp. 275–281, http://dx.doi.org/10.1145/290941.291008.

[60] C. Zhai, Statistical language models for information retrieval a critical review, Found. Trends® Inf. Retr. 2 (3) (2008) 137–213, http://dx.doi.org/10.1561/1500000008.

[61] Y. Lv, C. Zhai, A comparative study of methods for estimating query language models with pseudo feedback, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM '09, Association for Computing Machinery, New York, NY, USA, 2009, pp. 1895–1898, http://dx.doi.org/10.1145/1645953.1646259.

[62] V. Lavrenko, W.B. Croft, Relevance-based language models, in: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '01, Association for Computing Machinery, New York, NY, USA, 2001, pp. 120–127, http://dx.doi.org/10.1145/383952.383972.

[63] C. Zhai, J. Lafferty, Model-based feedback in the language modeling approach to information retrieval, in: Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01, Association for Computing Machinery, New York, NY, USA, 2001, pp. 403–410, http://dx.doi.org/10.1145/502585.502654.

[64] Y. Lv, C. Zhai, Revisiting the divergence minimization feedback model, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 1863–1866, http://dx.doi.org/10.1145/2661829.2661900.

[65] H. Hazimeh, C. Zhai, Axiomatic analysis of smoothing methods in language models for pseudo-relevance feedback, in: Proceedings of the 2015 International Conference on the Theory of Information Retrieval, ICTIR '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 141–150, http://dx.doi.org/10.1145/2808194.2809471.

[66] S. Robertson, Understanding inverse document frequency: on theoretical arguments for IDF, J. Doc. 60 (5) (2004) 503–520, http://dx.doi.org/10.1108/00220410410560582.

[67] D.E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019 early risk prediction on the internet, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D.E. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2019, pp. 340–357.

[68] D.E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2020: Early risk prediction on the internet, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2020, pp. 272–287.

[69] J. Parapar, P. Martín-Rodilla, D.E. Losada, F. Crestani, Overview of eRisk 2021: Early risk prediction on the internet, in: K.S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2021, pp. 324–344.

[70] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. (TOIS) 20 (4) (2002) 422–446.

[71] S. Nolen-Hoeksema, The role of rumination in depressive disorders and mixed anxiety/depressive symptoms, J. Abnorm. Psychol. 109 (3) (2000) 504.

[72] Y. Liu, L. Biester, R. Mihalcea, Improving mental health classifier generalization with pre-diagnosis data, Proc. Int. AAAI Conf. Web Social Media 17 (1) (2023) 566–577, http://dx.doi.org/10.1609/icwsm.v17i1.22169.

[73] K. Harrigian, C. Aguirre, M. Dredze, Do models of mental health based on social media data generalize? in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 3774–3788, http://dx.doi.org/10.18653/v1/2020.findings-emnlp.337, Online.

[74] K. Gligorić, A. Anderson, R. West, How constraints affect content: The case of Twitter's switch from 140 to 280 characters, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, (1) 2018.
[75] S.T. Ibrahim, M. Li, J. Patel, T.R. Katapally, Utilizing natural language processing for precision prevention of mental health disorders among youth: A systematic review, Comput. Biol. Med. 188 (2025) 109859.
[76] D.E. Losada, P. Gamallo, Evaluating and improving lexical resources for detecting signs of depression in text, Lang. Resour. Eval. 54 (1) (2020) 1–24, http://dx.doi.org/10.1007/s10579-018-9423-1.
[77] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 47–56, http://dx.doi.org/10.1145/2464464.2464480.
[78] J.W. Pennebaker, R.J. Booth, M.E. Francis, Linguistic Inquiry and Word Count: LIWC 2007, LIWC.net, 2007.
[79] S.M. Mohammad, P.D. Turney, NRC Emotion Lexicon, Tech. Rep., National Research Council of Canada, 2013, p. 234, http://dx.doi.org/10.4224/21270984, Collection / Collection : NRC Publications Archive / Archives des publications du CNRC / Record identifier / Identificateur de l'enregistrement : 0b6a5b58-a656-49d3-ab3e-252050a7a88c.
[80] European Commission, The digital services act: protecting users' fundamental rights and mental wellbeing, 2023, European Commission Publications, Retrieved from relevant EU publications.
[81] R. Blumenthal, Kids Online Safety Act (KOSA), United States Senate, https://www.blumenthal.senate.gov/about/issues/kids-online-safety-act,
[82] Meta, Mental Health & Well-Being, Meta Publications, Retrieved from Meta's official publications,
[83] Reddit Help, What do I do if someone talks about seriously hurting themselves or is considering suicide?, Reddit Help Center https://support.reddithelp.com/hc/en-us/articles/360043513931-What-do-I-do-if-someone-talks-about-seriously-hurting-themselves-or-is-considering-suicide.
[84] X Help Center, Suicide and Self-harm policy, X Help Center, https://help.x.com/en/rules-and-policies/glorifying-self-harm,