Original Paper

Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening: Instrument Validation Study

Daun Shin[1,2], MD, PhD; Hyoseung Kim[3], BS; Seunghwan Lee[3], BS; Younhee Cho[2,4], DA; Whanbo Jung[2], MD

[1]Department of Psychiatry, Anam Hospital, Korea University, Seoul, Republic of Korea

[2]Doctorpresso, Seoul, Republic of Korea

[3]VOLTWIN, Seoul, Republic of Korea

[4]Department of Design, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Daun Shin, MD, PhD
Department of Psychiatry
Anam Hospital
Korea University
73 Goryeodae-ro
Seongbuk-gu
Seoul, 02841
Republic of Korea
Phone: 82 1093649735
Fax: 82 2 920 5185
Email: rune1018@gmail.com

Related Article:
This is a corrected version. See correction statement in: https://www.jmir.org/2025/1/e79198

Abstract

Background: Depressive disorders have substantial global implications, leading to various social consequences, including decreased occupational productivity and a high disability burden. Early detection and intervention for clinically significant depression have gained attention; however, the existing depression screening tools, such as the Center for Epidemiologic Studies Depression Scale, have limitations in objectivity and accuracy. Therefore, researchers are identifying objective indicators of depression, including image analysis, blood biomarkers, and ecological momentary assessments (EMAs). Among EMAs, user-generated text data, particularly from diary writing, have emerged as a clinically significant and analyzable source for detecting or diagnosing depression, leveraging advancements in large language models such as ChatGPT.

Objective: We aimed to detect depression based on user-generated diary text through an emotional diary writing app using a large language model (LLM). We aimed to validate the value of the semistructured diary text data as an EMA data source.

Methods: Participants were assessed for depression using the Patient Health Questionnaire and suicide risk was evaluated using the Beck Scale for Suicide Ideation before starting and after completing the 2-week diary writing period. The text data from the daily diaries were also used in the analysis. The performance of leading LLMs, such as ChatGPT with GPT-3.5 and GPT-4, was assessed with and without GPT-3.5 fine-tuning on the training data set. The model performance comparison involved the use of chain-of-thought and zero-shot prompting to analyze the text structure and content.

Results: We used 428 diaries from 91 participants; GPT-3.5 fine-tuning demonstrated superior performance in depression detection, achieving an accuracy of 0.902 and a specificity of 0.955. However, the balanced accuracy was the highest (0.844) for GPT-3.5 without fine-tuning and prompt techniques; it displayed a recall of 0.929.

Conclusions: Both GPT-3.5 and GPT-4.0 demonstrated relatively reasonable performance in recognizing the risk of depression based on diaries. Our findings highlight the potential clinical usefulness of user-generated text data for detecting depression. In addition to measurable indicators, such as step count and physical activity, future research should increasingly emphasize qualitative digital expression.

(J Med Internet Res 2024;26:e54617) doi: 10.2196/54617

KEYWORDS

depression; screening; artificial intelligence; digital health technology; text data

Introduction

Depressive disorders are globally prevalent mental health conditions that significantly impact social and occupational functioning [1-3]. Major depressive disorder (MDD) and dysthymia are particularly noteworthy, as together, they were the second leading cause of years lived with disability in 2010, with MDD and dysthymia contributing 8.2% and 1.4%, respectively. The global incidence of depression increased from 172 million in 1990 to 258 million in 2017, reflecting a 49.86% rise, and the associated burden increased by 37.5% from 1990 to 2010 [4,5]. Moreover, depression is a key contributor to increased mortality risk, as evidenced by a meta-analysis indicating a hierarchy in the lifetime prevalence of suicide among patients with affective disorders [6,7]. Early detection and intervention for clinically significant depression have garnered increased attention. Prior to the 2000s, there was insufficient evidence supporting the use of screening tools. However, the United States Preventive Services Task Force revised its stance in June 2002, recommending that physicians screen for MDD [8].

Depressive symptoms are typically evaluated through self-reports or clinical assessment, with notable assessment methods including the Center for Epidemiologic Studies Depression (CES-D) scale, Patient Health Questionnaire–9 (PHQ-9), and Beck Depression Inventory [9-13]. However, these methods have limitations. Patients may underreport symptoms, and there can be discrepancies between subjective reports and objective severity [14,15]. Additionally, individuals often seek initial treatment from general practitioners rather than psychiatry specialists, partly due to stigma and lack of awareness, further complicating accurate assessment [16].

To address these limitations, researchers are exploring various biomarkers, genetic markers, and ecological momentary assessments (EMAs) for more objective and accurate screening [17-19]. EMAs involve real-time data collection, either actively by user input or passively through sensors on wearable devices [20]. Quantitative data such as exercise levels, step counts, and sleep cycles are relatively straightforward to collect and analyze in relation to depression scales [21]. However, analyzing qualitative data, such as user-generated text, presents a more complex challenge.

User-generated text data hold significant clinical potential. Advances in artificial intelligence (AI), particularly in natural language processing (NLP), have enabled sophisticated analysis of such data [22,23]. Large language models (LLMs) like ChatGPT (OpenAI) have facilitated various medical applications, including in psychiatry [24,25]. These models can analyze language used by patients in everyday contexts, such as social media posts, speech, or writing, to detect markers associated with depression [26,27]. By examining linguistic patterns, NLP and LLMs can predict depression risk without relying on traditional survey participation [28].

Despite their potential, much of the text data used in depression research has been sourced from electronic medical records or social media, which may not fully represent natural language use [29]. Daily writing is a universal human activity with therapeutic benefits, including improvements in depression and nonsuicidal self-injury [30-34]. Diary writing supports patient introspection, growth, and communication with therapists, making it a promising EMA data source.

In this study, we aimed to develop an algorithm for depression screening based on user-generated diary text data. Using a daily writing app for emotions, we collected semistructured diary texts and analyzed them with an LLM. Our goal was to validate the clinical utility of these texts as an EMA data source for identifying depression risk.

By leveraging the therapeutic and diagnostic potential of daily writing, combined with the analytical power of LLMs, we seek to contribute to the field of digital mental health. Our approach addresses the limitations of traditional screening tools and highlights the importance of qualitative data in mental health assessment and intervention.

Methods

App Process and Diary Log

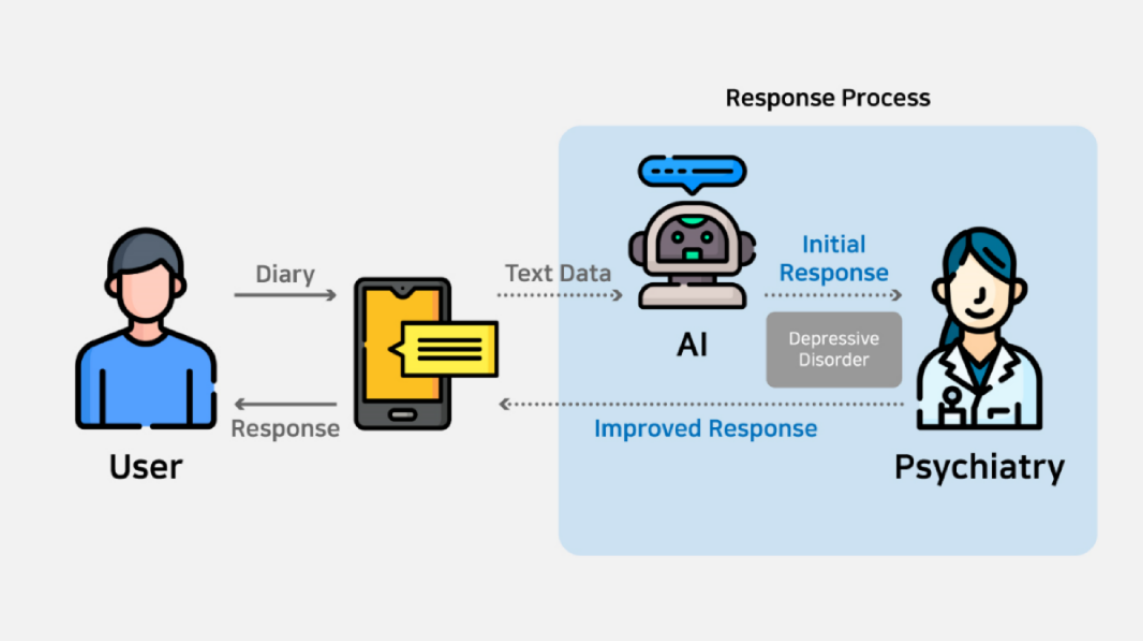
The text data were obtained through an app named Mind Station. This app prompts users to write daily emotional diaries. Based on diary data, an AI system assesses the risk of depression and provides this information to mental health professionals. In addition, it uses generative AI to create responses to emotional diaries. A psychiatric clinician reviews these responses to modify and enhance them in a cognitive-behavioral or supportive psychotherapeutic manner, then provides replies to the user's diary (Figure 1).

The daily diary log was divided into 4 paragraphs, beginning with a description of the events that occurred that day, followed by reflections on these events and the resulting emotions, and concluding with a free-form diary entry. For each paragraph, the limit was set to a maximum of 500 characters to ensure brevity. Based on the written diaries, a specialist in mental health and medicine provided individualized emotional support. This step involved encouraging exploration, rectifying distorted cognition, guiding antidepressive activities in daily life, and recommending medical assistance (if necessary). Our objective was to enhance mental health through numerous responses tailored to individual needs.

Furthermore, this app was not developed for research purposes but rather as a form of mental self-care. However, we had plans to collect various digital phenotype data through the Mind Station app and to build algorithms for diagnosing and treating depression by analyzing text data. Therefore, users of the app

were allowed to write diaries and receive responses from clinical psychiatrists only after agreeing that their anonymized data could be used for future research purposes.

Figure 1. Application process. AI: artificial intelligence.



Recruitment

We recruited 91 participants from October 1, 2022, to April 30, 2023. Participants were recruited through promotion within internet communities for beta testing of the app. These promotions informed individuals about an app for writing diaries to care for their mental health where they could receive free responses from clinical psychiatrists. Additionally, they were notified that their anonymized data might be used for future research purposes. The participants received a reward of approximately ₩30,000 (US $22) for completing diaries for at least 1 week, writing daily, and submitting self-reported questionnaires about their mood. Furthermore, they voluntarily agreed to receive information on data collection and use before using the app. In addition, rewards were only disbursed after the completion of the week-long app experience and the submission of feedback regarding technical errors or diary responses encountered in the app. This was done to reduce the likelihood of participants writing diaries arbitrarily. Researchers should collect well-established risk factors, such as sex, family history, and employment status, for depression screening. However, an increase in the amount of input information can lead to lower compliance, potentially diminishing its role as a screening tool. Therefore, we performed an analysis based solely on text data from diaries written by users and from self-reported questionnaires. Diaries that were filled only with words lacking substantive content, such as “lol,” or solely with names of other individuals who were present at a particular location, were excluded from the analysis. This study aimed to retrospectively analyze the collected data.

Ethical Considerations

This study adhered to the ethical principles outlined in the Helsinki Declaration. Approval was obtained from the Institutional Review Board of Korea University Anam Hospital (2023AN0379). Each participant received approximately ₩30,000 (US $22) for completing diaries as long as they fulfilled the criteria described above. Data were anonymized. The participants were informed and consented to the use of their data for research purposes upon accessing the app. As this study involved retrospective analysis of the data, no written informed consent was obtained.

Depression Assessment Scale, Diary Classification, and Statistics

All participants were evaluated for depressive symptoms using the PHQ-9 and for potential crises in suicide situations using the Beck Scale for Suicide Ideation (BSS). The PHQ-9 consists of 9 questions related to mood, sleep, and other factors, with scores ranging from 0 to 3 for each question. The total score ranges from 0 to 27 [35]. The optimal cutoff score for depression screening is 10, with scores ≥11 indicating a risk of depression [36]. The BSS is a self-reported questionnaire for assessing suicide risk, and its potential as a screening tool in emergency rooms and inpatient settings has been validated [37-39]. These scores are well established in clinical research, with the PHQ-9 widely recognized for its validity and reliability in assessing depressive symptoms.

The classification of diaries as depressive was based on a combination of the validated PHQ-9, the BSS, and the clinical psychiatrist’s review. The primary criterion for determining a “true” classification was the PHQ-9 and BSS scores. A participant was classified as “depressed” if their PHQ-9 score was 10 or higher or if their BSS score was 8 or higher at the closest point to the diary-writing day. For diaries without available PHQ-9 and BSS scores, a psychiatrist reviewed the content. Diaries were classified as depressive if they contained direct expressions of depression, such as “I want to die” or “I am depressed,” or if they had clinically recognizable signs of

depression, like "I am very stressed." Additionally, if a participant's PHQ-9 score was 10 or higher at the time of writing the first diary entry, but subsequent diaries (4-5 days later) included statements indicating an improved mood, such as "I feel better" or "Today was a good day," the psychiatrist reviewed these diaries and classified them as not exhibiting signs of depression. This dual approach of using validated depression scales and clinician review ensured a comprehensive and accurate classification of the diaries, balancing quantitative and qualitative assessments.

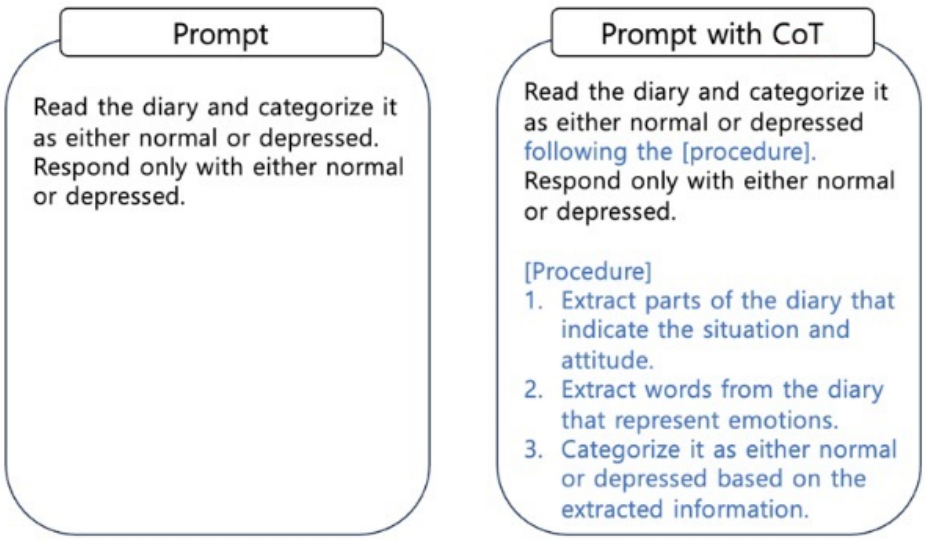
The statistical analysis method used for comparing depression and suicide scores before and after diary writing was a 2-tailed Student *t* test. Statistical analyses were conducted using SPSS Statistics (version 24.0; IBM Corp).

AI Technology and Data Preprocessing

LLMs have achieved remarkable success in the field of NLP. Initially, we applied the most popular LLMs, such as ChatGPT with GPT-3.5 and GPT-4 [40], to our data set without fine-tuning. In addition, we used a GPT-3.5 fine-tuning model with our training data set. Furthermore, we compared the performance of each model by applying chain-of-thought (CoT) and zero-shot prompting (Figure 2) [41,42]. CoT prompting, zero-shot prompting, and few-shot prompting were defined as follows in this study: CoT prompting is a method used in NLP to guide the model in generating intermediate reasoning steps leading to the final answer. Instead of directly producing an answer, the model generates a sequence of logical thoughts or steps connecting the input to the output. Zero-shot prompting refers to a scenario where an LLM is presented with a task without any prior specific examples or training on that task. The model relies on its preexisting knowledge and understanding to generate an appropriate response based on the provided prompt. Few-shot prompting is a technique in NLP where a model is given a small number of examples (usually ranging from one to a few dozen) of a task to learn from before being asked to perform the task itself. These examples aid the model in better understanding the task and improving its performance. Using the CoT approach, we structured the text and analyzed the words, expressions, and emotions used to classify it. Furthermore, we incorporated additional methods to compare the similarity of familiar texts, such as a compression algorithm based on the GNU Zip (Gzip) algorithm and k-nearest clustering. This step helped us compare the differences between the LLMs and conventional algorithms [43,44]. We adopted stratified 5-fold cross-validation to classify diaries while preserving the percentage of the ratio of labels. We divided the entire data set into training and test data sets at a ratio of 8:2.

In this study, we defined and calculated key metrics to evaluate the performance of our classification models. Recall, formerly known as sensitivity, is the proportion of actual positives correctly identified by the model. It is calculated by dividing the number of true positive predictions by the sum of true positive and false negative predictions. This metric helps in understanding how well the model can identify positive instances. Precision is the proportion of true positive predictions among the total positive predictions. It is calculated by dividing the number of true positive predictions by the sum of true positive and false positive predictions. Precision indicates the accuracy of the model's positive predictions. Specificity is the proportion of actual negatives correctly identified by the model. It is calculated by dividing the number of true negative predictions by the sum of true negative and false positive predictions. This metric is crucial for understanding the model's ability to correctly identify negative instances. Accuracy is the overall proportion of correct predictions, including both true positives and true negatives. It is calculated by dividing the sum of true positive and true negative predictions by the total number of predictions, which includes true positives, false positives, false negatives, and true negatives. Accuracy provides a general measure of how well the model performs across all classes [45].

Figure 2. Chain-of-thought (CoT) prompting. Differences with standard prompting are shown in blue.



Results

We collected 428 diaries from 91 participants, with an average of 4.7 (SD 4.44; median 7.0) diaries authored by each user. Among the participants, 34 consented to disclosing their sex and 31 consented to disclosing their age. At baseline, 85 participants responded to the PHQ-9, and at the end of the 2-week period, 34 participants responded. Similarly, 85 participants responded to the baseline BSS, and 30 participants responded at the end of the 2-week period. Among the respondents, 81% (25/31) were women, and 19% (6/31) were men. Regarding age, 87% (27/31) of the total participants were aged 20-39 years, while 13% (4/31) were aged 40-49 years. Every participant underwent an initial assessment using the PHQ-9 and BSS while writing their diaries. At baseline, 85 participants completed the PHQ-9. Their average score was 7.353 (SD 6.849), and 26 participants had scores of 10 or higher. Additionally, 85 participants completed the baseline BSS, and 30 participants completed it again at the end of the 2-week period. The average baseline BSS score was 4.200 (SD 5.708). Among the participants, 2 had PHQ-9 scores of 10 or lower and BSS scores of 8 or higher. As described in the Methods section, even if the initial PHQ-9 and BSS scores exceeded the cutoff points, diaries written 4-5 days later that included direct expressions of improved mood, such as “I feel better,” were classified as nondepressive diaries based on the psychiatrist’s judgment. Therefore, the total number of depressive diaries was 73. Furthermore, the PHQ-9 and BSS scores when terminating use of the app were lower than the scores at the beginning of use (PHQ-9: P=.32; BSS: P=.40) (Table 1).

Balanced accuracy was calculated as the average of the recall and specificity divided by 2, representing the accuracy for each class. This step is particularly useful for evaluating classes in unbalanced data sets. We used the F_1-score, accuracy, precision, recall, and specificity as evaluation metrics. GPT-3.5 fine-tuning demonstrated the best performance, with an accuracy of 0.902, F_1-score of 0.685, precision of 0.759, and specificity of 0.955. The best fold exhibited higher performance, with an accuracy of 0.942 and an F_1-score of 0.8. However, the recall for GPT-3.5 fine-tuning was moderate, at 0.643. By contrast, GPT-4 demonstrated the best performance, with a recall of 0.972. Nevertheless, GPT-4 lagged behind GPT-3.5 in fine-tuning regarding accuracy (0.743) and F_1-score (0.57) (Table 2). Figures 3 and 4 depict the combined confusion matrix for all folds of each model.

Table 1. Characteristics of participants.

Characteristics	Values
Sex of participants, n (%)	
Male	6 (19)
Female	25 (81)
Age group of participants (years), n (%)	
20-39	27 (87)
40-49	4 (13)
Patient Health Questionnaire–9 score , mean (SD)	
Baseline (n=85)	7.353 (6.849)
End (n=34)	5.735 (6.336)
Beck Scale for Suicide Ideation score , mean (SD)	
Baseline (n=85)	4.200 (5.708)
End (n=30)	2.967 (5.436)

Table 2. Performance of each algorithm with and without chain-of-thought (CoT) prompting. Italics indicate high performance.

Algorithm	Accuracy	Balanced accuracy	F_1-score	Precision	Recall	Specificity
Gpt3.5_ft						
Average	*0.883*	*0.808*[a]	0.670	0.653	0.695	*0.921*[b]
Maximum	0.930	0.853	0.786	0.846	0.733	0.972
Gpt3.5_ft_CoT						
Average	*0.902*[b]	0.799	0.685	0.759	0.643	*0.955*[c]
Maximum	0.942	0.833	0.800	1.000	0.667	1.000
Gpt3.5						
Average	0.789	*0.844*[a]	0.607	0.453	*0.929*[b]	0.761
Maximum	0.860	0.915	0.714	0.556	1.000	0.831
Gpt3.5_CoT						
Average	0.752	*0.818*[a]	0.560	0.405	*0.917*[b]	0.718
Maximum	0.802	0.828	0.605	0.464	0.867	0.789
Gpt4						
Average	0.713	*0.827*[a]	0.546	0.378	*1.000*[c]	0.653
Maximum	0.791	0.873	0.625	0.455	1.000	0.746
Gpt4_CoT						
Average	0.743	*0.834*[a]	0.570	0.406	*0.972*[c]	0.696
Maximum	0.826	0.894	0.667	0.500	1.000	0.789
Compression						
Average	*0.846*[a]	0.651	0.431	0.584	0.356	*0.947*[b]
Maximum	0.919	0.819	0.741	0.833	0.667	0.972

[a]Performance >0.80.

[b]Performance >0.90.

[c]Performance >0.95.

Figure 3. Confusion matrix for GPT 3.5 models. CoT: chain of thought; LLM: large language model.

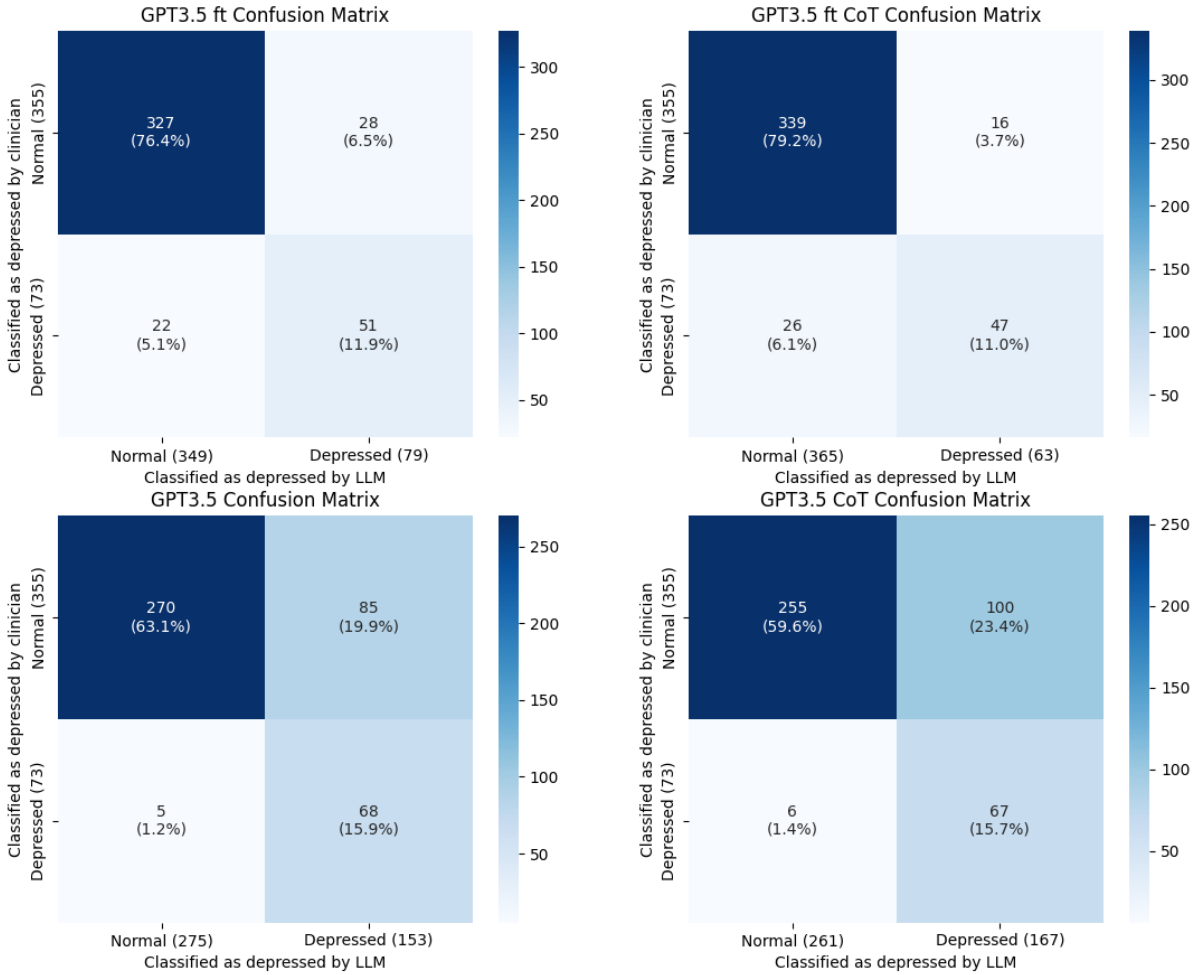
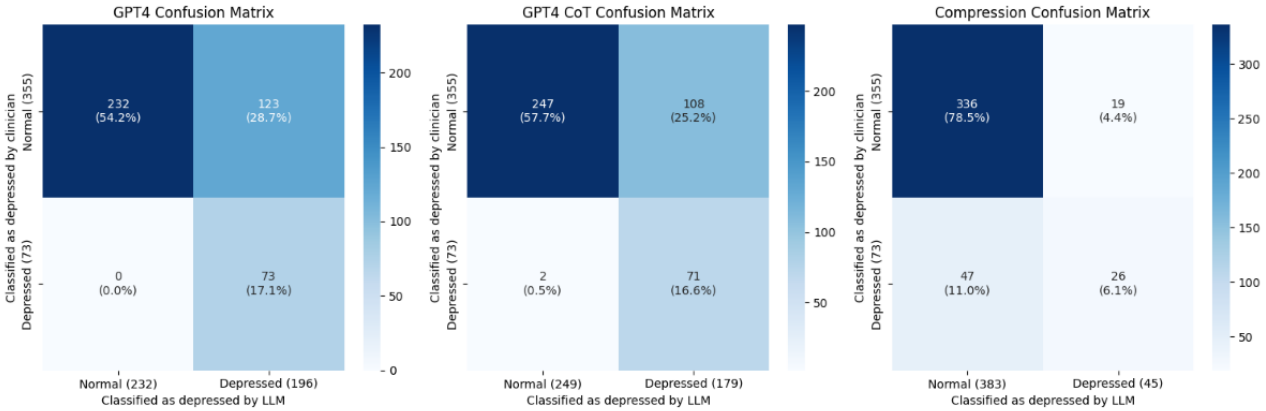


Figure 4. Confusion matrix for other artificial intelligence models. CoT: chain of thought.



Discussion

Principal Findings

In this study, log data comprising 428 daily diaries were collected from 91 participants. We predicted the risk of depression based solely on text data from the diaries, excluding other clinical data. Approximately 80% of the users were women, and the majority were aged 20-39 years. The participants were prompted to write diaries about their moods and daily experiences. The average initial PHQ-9 score was 5.719, indicating most users did not exhibit a significant risk of depressive symptoms. The GPT-3.5 model, combined with CoT prompting, achieved an accuracy of 0.902 and a specificity of 0.955 in classifying high-risk cases of depression based on diary text data. GPT-4.0 showed superior recall (1.000) compared to the GPT-4.0 CoT model (0.972), although both models demonstrated similar balanced accuracy. The high imbalance in the data set, with 82.9% (355/428) nondepressive and 17.1% (73/428) depressive diaries, underscores the importance of using balanced accuracy as a metric.

The analysis reveals that while the GPT-3.5 fine-tuned model had high accuracy and specificity, suggesting effectiveness in predicting false outcomes, its relatively low balanced accuracy implies potential overfitting during the fine-tuning process and inadequate prediction of true outcomes. On the other hand, GPT-4 achieved perfect recall but may have overlooked errors in false predictions due to its focus on distinguishing between normal and abnormal data. This raises concerns about false predictions with the GPT-4 model. Despite GPT-4's lower performance compared to GPT-3.5, both models had balanced accuracy metrics that remained similar. The similarity in performance can be attributed to factors such as the sentiment task nature of the query and the reinforcement of ethical filters in GPT-4. These findings align with the existing literature and highlight the need for further investigation into AI model performance and ethical considerations in sentiment analysis tasks.

Language, particularly text data, excluding voice, is used by individuals and reflects their values and emotions. It can be utilized for EMA. However, various methodological limitations make it challenging to use EMA for depression screening. Previous research on diagnosing or detecting depressive symptoms using text has been based on data from various sources, including time-based narratives, suicide notes, and publicly available social media content. These studies aimed to detect depression not by analyzing the text itself but by extracting various indicators, such as the counts of positive and negative words. These indicators were used as proxies for depression detection [46-48]. Furthermore, social media posts and poems often contain abbreviations or words that are not typically used in everyday, natural-language conversations. Thus, using such data as effective digital biomarkers poses limitations because they may not represent real-life communication accurately. With the evolution of various text analysis technologies, such as NLP, the scope of text used for screening and diagnosing depression has expanded to include electronic medical records [49,50]. However, the language used in electronic medical records is often influenced by the patient's clinical judgment, which can result in an altered representation. Consequently, text-based depression screening based on electronic medical record data has limitations and poor generalizability.

Fine-tuning is the process of optimizing a pretrained model through additional training for a specific domain or task. Concentrated learning of data related to a specific domain or task can improve performance while conducting tasks related to that domain. CoT is a simple and popular prompting algorithm for improving the performance of language models without fine-tuning. Our results demonstrate the performance of GPT-3.5 and GPT-4 on classification tasks and their improvement upon techniques for additional prompting and fine-tuning. The compression algorithm Gzip is an easy and lightweight nonparametric alternative to neural networks used for text classification. Gzip demonstrates good performance even in out-of-distribution data sets. We determined the results that would emerge from using existing technology instead of deep learning.

Mood diaries exert therapeutic effects. Our results suggest lower depression and suicide scores after completing diary writing than scores after commencing; nonetheless, the difference was statistically insignificant. This finding is partly attributable to the insufficient sample size to prove statistical power. Moreover, the study was conducted in the general population, and few participants demonstrated significantly higher levels of depressive symptoms. The therapeutic effects of diary writing on depression should be confirmed through large-scale studies with a larger sample size.

This study confirmed that emotion diaries among the general population can be used to screen for depression at a level similar to that of passive EMA. The area under the receiver operating characteristic curve of depression through passive sensing depends on the data used; however, the accuracy ranges from approximately 60% to 90% [51]. Text-based depression prediction has a lower accuracy (approximately 0.632); nonetheless, the accuracy of depression diagnosis increases upon using text data from medical interviews along with other data, such as vocal features [52]. Accuracy increases upon analyzing a daily mood diary rather than predicting depression and suicidal thoughts through EMA data obtained through a wearable device [53]. This study was conducted in a relatively large population; however, it has the disadvantage of poor generalizability because the targets were medical interns. A previous study suggested that journaling by older adults can help detect depression, and through this study, we confirmed that depression detection based on journaling can be extended to other age groups [54].

The findings of this research have important clinical implications. If depression could be diagnosed based on daily diaries, it would offer significant clinical benefits. Early detection could identify at-risk individuals, allowing for timely interventions like counseling or therapy to prevent symptom progression. Analyzing diary text data could reveal triggers and patterns, informing personalized therapy approaches such as cognitive behavioral therapy to address negative thought patterns or environmental factors. Regular diary analysis could monitor treatment effectiveness, providing real-time feedback for adjustments and improving patient outcomes. In summary, leveraging daily diaries for depression diagnosis could revolutionize clinical practice by enabling early intervention, personalized therapy, and ongoing treatment monitoring.

Strengths

The strength of this study is that we confirmed the basis for predicting mild levels of depression. This is because the recruitment of participants was not centered on hospitals. Most text data, excluding social media data, are based on medical records. Therefore, the data are supposedly biased toward individuals with severe depression. An existing depression diagnostic algorithm based on the frequency of negative language demonstrated low accuracy upon application to our data [52]. In addition, the text data are meaningful in that they can be used for simultaneous diagnosis and treatment because writing a diary is already therapeutic. Therefore, they can be a good evaluation tool for determining the severe burden placed on patients from obtaining active EMA data [55]. In addition,

we confirmed that prompt engineering alone affected depression detection. Thus, generative AI can facilitate medical judgment, and prompt technology is important.

Limitations

This study has several limitations. First, the complete exclusion of other clinical data, such as sex, age, and occupation, serves as both an advantage and a disadvantage. The advantage is that it confirms the diagnostic accuracy of pure text data; however, the sample size was small, and the predominant users were women aged 20-39 years. Therefore, the data are possibly biased. Second, we used generative AI; thus, we could not confirm the clinical characteristics recorded in the depression diary. Existing studies have demonstrated clinical implications, such as the use of *I*-centered words in suicide diaries and the frequent use of negative language in social media about depression [46,48]. In addition, the small sample size warrants large-scale research to diagnose depression using written diaries and the clinical characteristics of patients with depression. Third, text data may have a high possibility of being used in combination with other biomarkers; however, such data were not collected in this study. Therefore, further research is required on various types of EMA data that can be used in algorithms to differentiate depression in the general population.

Conclusions

We investigated whether an LLM can detect depression based on user-generated emotional diaries. Using a data set comprising 428 diaries from 91 users, the accuracy of predicting depression increased upon applying the CoT prompt technique to both GPT-3.5 and GPT-4.0. GPT-3.5 generated the highest average accuracy (0.902). However, without fine-tuning or prompting techniques, GPT-3.5 exhibited the highest balanced accuracy of 0.844 and recall of 0.929. Despite decreased depressive scores after the participants began writing their diaries, this change was statistically insignificant. Therefore, additional research is warranted to explore the potential antidepressant effects of diary writing. Based on these findings, we identified the potential clinical utility of voluntarily generated text data for detecting depression. Beyond quantifiable indicators, such as step count, daily physical activity, and sleep duration, we consider that there will be a continued and increased focus on qualitative digital expressions in research.

Acknowledgments

We express our gratitude to all the participants who contributed to this paper.

Conflicts of Interest

DS is the chief medical officer at Doctorpresso. While DS does not receive a salary from the company, she owns 25% of its equity. YC is a project manager at Doctorpresso and receives a salary for her work. WJ is the chief executive officer of Doctorpresso and owns 61% of the company's equity. HK and SL receive salaries from VOLTWIN but do not own any company equity. The Mind Station app was created by DS in collaboration with WJ and YC with the aim of providing mental health care based on writing emotion diaries. The app is owned by Doctorpresso. The data collected through this app were analyzed by VOLTWIN, a company specializing in data analysis. It is important to note that the 4 coauthors, aside from the first and corresponding author, DS, did not influence the study's design or analysis. YC contributed to the design of the Mind Station app, and WJ worked on community outreach projects based on the Mind Station app. VOLTWIN was subcontracted to perform artificial intelligence–related analysis, and the analysis plan was carried out in consultation with DS, ensuring that the results were not affected by the company. This work was also supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health & Welfare, Republic of Korea (grant HI23C0035). The funding source had no role in the study design, data collection, analysis, or interpretation or in the decision to submit the manuscript for publication.

References

1. Ferrari AJ, Charlson FJ, Norman RE, Flaxman AD, Patten SB, Vos T, et al. The epidemiological modelling of major depressive disorder: application for the Global Burden of Disease Study 2010. PLoS One. 2013;8(7):e69637. [FREE Full text] [doi: 10.1371/journal.pone.0069637] [Medline: 23922765]
2. Ferrari AJ, Somerville AJ, Baxter AJ, Norman R, Patten SB, Vos T, et al. Global variation in the prevalence and incidence of major depressive disorder: a systematic review of the epidemiological literature. Psychol Med. Mar 2013;43(3):471-481. [doi: 10.1017/S0033291712001511] [Medline: 22831756]
3. Moreno-Agostino D, Wu Y, Daskalopoulou C, Hasan MT, Huisman M, Prina M. Global trends in the prevalence and incidence of depression:a systematic review and meta-analysis. J Affect Disord. Feb 15, 2021;281:235-243. [doi: 10.1016/j.jad.2020.12.035] [Medline: 33338841]
4. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJ, et al. Burden of depressive disorders by country, sex, age, and year: findings from the Global Burden of Disease Study 2010. PLoS Med. Nov 2013;10(11):e1001547. [FREE Full text] [doi: 10.1371/journal.pmed.1001547] [Medline: 24223526]
5. Liu Q, He H, Yang J, Feng X, Zhao F, Lyu J. Changes in the global burden of depression from 1990 to 2017: findings from the Global Burden of Disease study. J Psychiatr Res. Jul 2020;126:134-140. [FREE Full text] [doi: 10.1016/j.jpsychires.2019.08.002] [Medline: 31439359]
6. Lépine J-P, Briley M. The increasing burden of depression. Neuropsychiatr Dis Treat. 2011;7(Suppl 1):3-7. [FREE Full text] [doi: 10.2147/NDT.S19617] [Medline: 21750622]

7. Bostwick JM, Pankratz VS. Affective disorders and suicide risk: a reexamination. Am J Psychiatry. Dec 2000;157(12):1925-1932. [doi: 10.1176/appi.ajp.157.12.1925] [Medline: 11097952]
8. U.S. Preventive Services Task Force. Screening for depression: recommendations and rationale. Ann Intern Med. May 21, 2002;136(10):760-764. [FREE Full text] [doi: 10.7326/0003-4819-136-10-200205210-00012] [Medline: 12020145]
9. Radloff LS. The CES-D Scale. Appl Psychol Meas. Jul 26, 2016;1(3):385-401. [doi: 10.1177/014662167700100306]
10. Vieweg BW, Hedlund JL. The General Health Questionnaire (GHQ): a comprehensive review. J Op Psychiatry. 1983;14(2):74-81.
11. Ball K, MacPherson C, Hurowitz G, Settles-Reaves B, DeVeaugh-Geiss J, Weir S. M3 checklist and SF-12 correlation study. Best Pract Ment Health. Mar 2015;11(1):83-89(7). [FREE Full text] [doi: 10.4324/9781315795621-9]
12. Manea L, Gilbody S, McMillan D. A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. Gen Hosp Psychiatry. 2015;37(1):67-75. [doi: 10.1016/j.genhosppsych.2014.09.009] [Medline: 25439733]
13. Lasa L, Ayuso-Mateos J, Vázquez-Barquero JL, Díez-Manrique FJ, Dowrick C. The use of the Beck Depression Inventory to screen for depression in the general population: a preliminary analysis. J Affect Disord. 2000;57(1-3):261-265. [doi: 10.1016/s0165-0327(99)00088-9] [Medline: 10708841]
14. Hunt M, Auriemma J, Cashaw ACA. Self-report bias and underreporting of depression on the BDI-II. J Pers Assess. Feb 2003;80(1):26-30. [doi: 10.1207/s15327752jpa8001_10]
15. Ma S, Kang L, Guo X, Liu H, Yao L, Bai H, et al. Discrepancies between self-rated depression and observed depression severity: The effects of personality and dysfunctional attitudes. Gen Hosp Psychiatry. 2021;70:25-30. [doi: 10.1016/j.genhosppsych.2020.11.016] [Medline: 33689981]
16. Parslow RA, Jorm AF. Who uses mental health services in Australia? An analysis of data from the National Survey of Mental Health and Wellbeing. Aust N Z J Psychiatry. Dec 17, 2000;34(6):997-1008. [doi: 10.1046/j.1440-1614.2000.00839.x]
17. Le-Niculescu H, Kurian SM, Yehyawi N, Dike C, Patel SD, Edenberg HJ, et al. Identifying blood biomarkers for mood disorders using convergent functional genomics. Mol Psychiatry. Feb 2009;14(2):156-174. [doi: 10.1038/mp.2008.11] [Medline: 18301394]
18. Abi-Dargham A, Horga G. The search for imaging biomarkers in psychiatric disorders. Nat Med. Nov 2016;22(11):1248-1255. [doi: 10.1038/nm.4190] [Medline: 27783066]
19. Rykov Y, Thach T, Bojic I, Christopoulos G, Car J. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. JMIR Mhealth Uhealth. Oct 25, 2021;9(10):e24872. [FREE Full text] [doi: 10.2196/24872] [Medline: 34694233]
20. Kamath J, Leon Barriera R, Jain N, Keisari E, Wang B. Digital phenotyping in depression diagnostics: integrating psychiatric and engineering perspectives. World J Psychiatry. Mar 19, 2022;12(3):393-409. [FREE Full text] [doi: 10.5498/wjp.v12.i3.393] [Medline: 35433319]
21. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res. Jul 15, 2015;17(7):e175. [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
22. Cohen KB, Hunter L. Getting started in text mining. PLoS Comput Biol. Jan 2008;4(1):e20. [FREE Full text] [doi: 10.1371/journal.pcbi.0040020] [Medline: 18225946]
23. Rzhetsky A, Seringhaus M, Gerstein MB. Getting started in text mining: part two. PLoS Comput Biol. Jul 2009;5(7):e1000411. [FREE Full text] [doi: 10.1371/journal.pcbi.1000411] [Medline: 19649304]
24. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res. Jun 2016;25(2):86-100. [FREE Full text] [doi: 10.1002/mpr.1481] [Medline: 26184780]
25. Cheng S, Chang C-W, Chang W-J, Wang H-W, Liang C-S, Kishimoto T, et al. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci. Nov 2023;77(11):592-596. [FREE Full text] [doi: 10.1111/pcn.13588] [Medline: 37612880]
26. Babu NV, Kanaga EGM. Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN Comput Sci. 2022;3(1):74. [FREE Full text] [doi: 10.1007/s42979-021-00958-1] [Medline: 34816124]
27. Nanomi Arachchige IA, Sandanapitchai P, Weerasinghe R. Investigating machine learning and natural language processing techniques applied for predicting depression disorder from online support forums: a systematic literature review. Information. Oct 27, 2021;12(11):444. [doi: 10.3390/info12110444]
28. Tejaswini V, Sathya Babu K, Sahoo B. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. ACM Trans Asian Low-Resour Lang Inf Process. Jan 15, 2024;23(1):1-20. [doi: 10.1145/3569580]
29. Meng Y, Speier W, Ong MK, Arnold CW. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. IEEE J Biomed Health Inform. Aug 2021;25(8):3121-3129. [doi: 10.1109/jbhi.2021.3063721]
30. Bolger N, Davis A, Rafaeli E. Diary methods: capturing life as it is lived. Annu Rev Psychol. 2003;54(1):579-616. [doi: 10.1146/annurev.psych.54.101601.145030] [Medline: 12499517]

31. Chen Y, Ishak Z. Gratitude diary: the impact on depression symptoms. Psych. Mar 2022;13(03):443-453. [doi: 10.4236/PSYCH.2022.133030]
32. Hooley JM, Fox KR, Wang SB, Kwashie AND. Novel online daily diary interventions for nonsuicidal self-injury: a randomized controlled trial. BMC Psychiatry. Aug 22, 2018;18(1):264. [FREE Full text] [doi: 10.1186/s12888-018-1840-6] [Medline: 30134866]
33. Abdallah CG, Jackowski A, Salas R, Gupta S, Sato JR, Mao X, et al. The nucleus accumbens and ketamine treatment in major depressive disorder. Neuropsychopharmacology. Jul 2017;42(8):1739-1746. [FREE Full text] [doi: 10.1038/npp.2017.49] [Medline: 28272497]
34. Faccio E, Turco F, Iudici A. Self-writing as a tool for change: the effectiveness of a psychotherapy using diary. Res Psychother. Aug 09, 2019;22(2):378. [FREE Full text] [doi: 10.4081/ripppo.2019.378] [Medline: 32913803]
35. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. Sep 2001;16(9):606-613. [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
36. Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. Gen Hosp Psychiatry. 2015;37(6):567-576. [doi: 10.1016/j.genhosppsych.2015.06.012] [Medline: 26195347]
37. Beck AT, Kovacs M, Weissman A. Assessment of suicidal intention: the Scale for Suicide Ideation. J Consult Clin Psychol. Apr 1979;47(2):343-352. [doi: 10.1037//0022-006x.47.2.343] [Medline: 469082]
38. Healy DJ, Barry K, Blow F, Welsh D, Milner KK. Routine use of the Beck Scale for Suicide Ideation in a psychiatric emergency department. Gen Hosp Psychiatry. 2006;28(4):323-329. [doi: 10.1016/j.genhosppsych.2006.04.003] [Medline: 16814632]
39. Pinninti N, Steer R, Rissmiller D, Nelson S, Beck A. Use of the Beck Scale for Suicide Ideation with psychiatric inpatients diagnosed with schizophrenia, schizoaffective, or bipolar disorders. Behav Res Ther. Sep 2002;40(9):1071-1079. [doi: 10.1016/s0005-7967(02)00002-5] [Medline: 12296492]
40. GPT-4. OpenAI. URL: https://openai.com/research/gpt-4 [accessed 2024-09-13]
41. Wei J, Wang X, Schuurmans D, Bosma M. Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, editors. NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY. Curran Associates; 2022:24824-24837.
42. Brown T, Mann B, Ryder N. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, editors. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY. Curran Associates; 2020:1877-1901.
43. Jiang Z, Yang M, Tsirlin M. "Low-resource" text classification: a parameter-free classification method with compressors. In: Rogers A, Boyd-Graber J, Okazaki N, editors. Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA. Association for Computational Linguistics; 2023:6810-6828.
44. Salvador-Meneses J, Ruiz-Chavez Z, Garcia-Rodriguez J. Compressed NN: K-nearest neighbors with data compression. Entropy (Basel). Feb 28, 2019;21(3):234. [FREE Full text] [doi: 10.3390/e21030234] [Medline: 33266949]
45. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. Apr 08, 2022;12(1):5979. [FREE Full text] [doi: 10.1038/s41598-022-09954-8] [Medline: 35395867]
46. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci. Dec 2017;18:43-49. [doi: 10.1016/j.cobeha.2017.07.005]
47. Rude S, Gortner E, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cogn Emot. Dec 2004;18(8):1121-1133. [doi: 10.1080/02699930441000030]
48. Stirman SW, Pennebaker JW. Word use in the poetry of suicidal and nonsuicidal poets. Psychosom Med. 2001;63(4):517-522. [doi: 10.1097/00006842-200107000-00001] [Medline: 11485104]
49. Tsui FR, Shi L, Ruiz V, Ryan ND, Biernesser C, Iyengar S, et al. Natural language processing and machine learning of electronic health records for prediction of first-time suicide attempts. JAMIA Open. Jan 2021;4(1):ooab011. [FREE Full text] [doi: 10.1093/jamiaopen/ooab011] [Medline: 33758800]
50. Zhong Q, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, et al. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. BMC Med Inform Decis Mak. May 29, 2018;18(1):30. [doi: 10.1186/s12911-018-0617-7]
51. Colombo D, Fernández-Álvarez J, Patané A, Semonella M, Kwiatkowska M, García-Palacios A, et al. Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: a systematic review. J Clin Med. Apr 05, 2019;8(4):465. [FREE Full text] [doi: 10.3390/jcm8040465] [Medline: 30959828]
52. Shin D, Kim K, Lee S, Lee C, Bae YS, Cho WI, et al. Detection of depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker. Front Psychiatry. 2022;13:801301. [FREE Full text] [doi: 10.3389/fpsyt.2022.801301] [Medline: 35686182]
53. Horwitz A, Czyz E, Al-Dajani N, Dempsey W, Zhao Z, Nahum-Shani I, et al. Utilizing daily mood diaries and wearable sensor data to predict depression and suicidal ideation among medical interns. J Affect Disord. Sep 15, 2022;313:1-7. [FREE Full text] [doi: 10.1016/j.jad.2022.06.064] [Medline: 35764227]

54. Chepenik LG, Have TT, Oslin D, Datto C, Zubritsky C, Katz IR. A daily diary study of late-life depression. Am J Geriatr Psychiatry. Mar 2006;14(3):270-279. [doi: 10.1097/01.JGP.0000194644.63245.42] [Medline: 16505132]
55. Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, et al. Ecological momentary assessment in behavioral research: addressing technological and human participant challenges. J Med Internet Res. Mar 15, 2017;19(3):e77. [FREE Full text] [doi: 10.2196/jmir.7138] [Medline: 28298264]

Abbreviations

AI: artificial intelligence
BSS: Beck Scale for Suicide Ideation
CES-D: Center For Epidemiologic Studies Depression
CoT: chain-of-thought
EMA: ecological momentary assessment
Gzip: GNU Zip
LLM: large language model
MDD: major depressive disorder
NLP: natural language processing
PHQ-9: Patient Health Questionnaire–9

Edited by G Eysenbach; submitted 16.11.23; peer-reviewed by Z Su, L Luo; comments to author 30.04.24; revised version received 17.05.24; accepted 11.08.24; published 18.09.24

Please cite as:
Shin D, Kim H, Lee S, Cho Y, Jung W
Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening: Instrument Validation Study
J Med Internet Res 2024;26:e54617
URL: https://www.jmir.org/2024/1/e54617
doi: 10.2196/54617
PMID: 39292502