Natural Language Processing and Social Determinants of Health in Mental Health Research: An Artificial Intelligence Assisted Scoping Review

Dmitry Scherbakov, Nina Hubig, Leslie Andrew Lenert, Alexander V. Alekseyenko, Jihad S. Obeid

Submitted to: JMIR Mental Health
on: October 04, 2024

Table of Contents

Natural Language Processing and Social Determinants of Health in Mental Health Research: An Artificial Intelligence Assisted Scoping Review

Dmitry Scherbakov[1] PhD; Nina Hubig[1] PhD; Leslie Andrew Lenert[1] MD, MS, FACP; Alexander V. Alekseyenko[1] PhD, FAMIA; Jihad S. Obeid[1] MD

[1]Biomedical Informatics Center Medical University of South Carolina Charleston US

Corresponding Author:
Jihad S. Obeid MD
Biomedical Informatics Center
Medical University of South Carolina
22 WestEdge St.
Charleston
US

Abstract

Background: The usage of natural language processing (NLP) in mental health research is increasing with a wide range of applications and datasets being investigated.

Objective: This review aims to summarize the usage NLP in mental health research, with a special focus on the types of text datasets and the usage of social determinants of health (SDOH) in NLP projects related to mental health.

Methods: The search was conducted in September 2024 using a broad search strategy in PubMed, Scopus, and CINAHL Complete. All citations were uploaded to Covidence online software. The screening and extraction process took place in Covidence with the help of a custom large language model (LLM) module developed by our team. This LLM module was calibrated and tuned to substitute human reviewers.

Results: The screening process, assisted by the custom LLM, led to the inclusion of 1,768 studies in the final review. The majority of the reviewed studies (n=665, 42.8%) utilized clinical data as their primary text dataset, followed by social media datasets (n=523, 33.7%). The United States contributed the highest number of studies (n=568, 36.6%), with depression (n=438, 28.2%) and suicide (n=240, 15.5%) being the most frequently investigated mental health issues. Traditional demographic variables such as age (n=877, 56.5%) and gender (n=760, 49.0%) were commonly extracted, while SDOH factors were less frequently reported, with urban/rural status being the most used (n=19, 1.2%). Over half of the citations (n=826, 53.2%) did not provide clear information on dataset accessibility, although a sizable number of studies (n=304, 19.6%) made their datasets publicly available.

Conclusions: This scoping review underscores the significant role of clinical notes and social media in NLP-based mental health research. Despite the clear relevance of SDOH to mental health, their underutilization presents a gap in current research. This review can be a starting point for researchers looking for an overview of mental health projects using text data. Discovered datasets could be used to place more emphasis on SDOH in future studies.

(JMIR Preprints 04/10/2024:67192)
DOI: https://doi.org/10.2196/preprints.67192

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Natural Language Processing and Social Determinants of Health in Mental Health Research: An Artificial Intelligence Assisted Scoping Review

Dmitry Scherbakov, PhD[1], Nina Hubig, PhD[1,2], Leslie A. Lenert, MD[1], Alexander V. Alekseyenko, PhD[1], Jihad S. Obeid, MD[1]

[1] Biomedical Informatics Center, Department of Public Health Sciences, Medical University of South Carolina (MUSC), Charleston, South Carolina, USA.

[2] Clemson University, School of Computing, Charleston, South Carolina, USA.

Corresponding author: Jihad S. Obeid (jobeid@musc.edu)

Abstract

Background: The usage of natural language processing (NLP) in mental health research is increasing with a wide range of applications and datasets being investigated.

Objective: This review aims to summarize the usage NLP in mental health research, with a special focus on the types of text datasets and the usage of social determinants of health (SDOH) in NLP projects related to mental health.

Methods: The search was conducted in September 2024 using a broad search strategy in PubMed, Scopus, and CINAHL Complete. All citations were uploaded to Covidence online software. The screening and extraction process took place in Covidence with the help of a custom large language model (LLM) module developed by our team. This LLM module was calibrated and tuned to substitute human reviewers.

Results: The screening process, assisted by the custom LLM, led to the inclusion of 1,768 studies in the final review. The majority of the reviewed studies (n=665, 42.8%) utilized clinical data as their primary text dataset, followed by social media datasets (n=523, 33.7%). The United States contributed the highest number of studies (n=568, 36.6%), with depression (n=438, 28.2%) and suicide (n=240, 15.5%) being the most frequently investigated mental health issues. Traditional demographic variables such as age (n=877, 56.5%) and gender (n=760, 49.0%) were commonly extracted, while SDOH factors were less frequently reported, with urban/rural status being the most used (n=19, 1.2%). Over half of the citations (n=826, 53.2%) did not provide clear information on

dataset accessibility, although a sizable number of studies (n=304, 19.6%) made their datasets publicly available.

Conclusions: This scoping review underscores the significant role of clinical notes and social media in NLP-based mental health research. Despite the clear relevance of SDOH to mental health, their underutilization presents a gap in current research. This review can be a starting point for researchers looking for an overview of mental health projects using text data. Discovered datasets could be used to place more emphasis on SDOH in future studies.

Keywords: Natural Language Processing, Datasets, Mental Health, Automated review.

Introduction

Natural Language Processing (NLP) has emerged as a valuable tool in mental health research, offering innovative ways to extract and analyze information from various sources. Studies have shown the feasibility of using NLP in extracting evidence of online activity from electronic health records (EHRs) in adolescent mental health patients [1]. NLP applied to clinical notes has been found to more accurately identify mental illness and substance use among people living with HIV compared to structured EHR fields alone [2]. Furthermore, NLP in healthcare enables the transformation of complex narrative information into valuable products like clinical decision support and adverse event monitoring in real-time via EHRs [3, 4].

Outside of EHRs, NLP techniques have been utilized to make inferences about individuals' mental states based on their social media posts [5]. Additionally, NLP, coupled with machine learning approaches, has shown promising performance in tasks such as text classification and sentiment mining in mental health contexts [6]. The application of NLP extends to identifying work-related stress among health professionals, highlighting its versatility in diverse healthcare settings [7].

In the context of mental health disorders like schizophrenia, schizoaffective disorder, and bipolar disorder, NLP applied to EHRs offers opportunities to create large datasets for research purposes [8]. Furthermore, NLP has been employed to increase prediction accuracy and reduce subgroup differences in personnel selection decisions, showcasing its value in improving decision-making processes [9].

At the same time, getting access to text datasets for NLP analysis is a challenging task for many researchers. Many of the datasets have strict privacy and personal data protection policies restricting access to the data for 3rd party researchers. This hinders the research and introduces the

problem of reproducibility since the results of the studies cannot be verified by unaffiliated investigators. One of the aims of this review is to compile a collection of datasets that researchers use. This information may facilitate the research in the mental health area.

Another potential problem with research using NLP for mental health is insufficient consideration of social determinants of health (SDOH) information during the analysis. The association between social determinants and mental health outcomes is well-established, with factors such as poverty, inequality, stigma, discrimination, and social exclusion identified as significant contributors to mental health burdens [10, 11]. NLP has become a valuable tool for extracting SDOH from sources like clinical notes, social media, and EHR in healthcare research [12-15]. However, the pilot review of mental health projects using NLP methods revealed significant gaps in this area: social determinants are seldomly considered, as studies most often focus on basic demographic information, such as age and gender. Thus, evaluating the usage of social determinants of health in NLP projects for mental health is another goal behind this review.

To our knowledge, no previous study has examined the range of NLP datasets and the usage of SDOH data in research projects that use NLP for mental health. We have opted for the scoping review following the guidelines outlined by Arksey & O'Malley [16]. The goals of this scoping review is to review and summarize the literature on (1) the natural language processing projects in different mental health areas, (2) information on the text datasets used in these projects, (3) whether and which social determinants of health information were used in these projects.

This review's novelty lies in using large language models (LLMs) to automatically parse a large volume of citations to find relevant studies and extract information under the minimal supervision of a human reviewer. A recent statement by the National Institute for Health and Care Excellence (NICE) highlights the potential of artificial intelligence (AI) in the systematic review process automation [17].

Methods

This study was created and revised following the recommendation of Preferred Reporting Items for Systematic Reviews and Meta-analysis Protocols extension for Scoping Reviews (PRISMA-ScR) and updated JBI (formerly known as Joanna Briggs Institute) guidance for the conduct of scoping reviews [16, 18-20]. The completed PRISMA-ScR checklist can be found in Supplementary Appendix S1.

All publications were considered if they did not meet one or more of the exclusion criteria. Citations were excluded if they:

- Did not utilize some kind of natural language processing method (NLP), like transformers, pattern-matching (e.g., regular expressions), ChatGPT, GPT-3, BERT, Llama, Mistral, large language models (LLM), Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), deep learning or machine learning applied to text, and similar.
- Were not focused on one of the mental health areas, such as psychology, well-being, psychiatry, social work, substance abuse, marriage therapy, addiction therapy, suicide, grief, bereavement, trauma, stressful life events, counseling,
- Were review papers (systematic, scoping, literature, narrative, and other type of reviews), conference papers, book chapters
- Were not related to human health or well-being
- Portable Document Format (PDF) file of the full-text publication could not be located automatically using Covidence, EndNote, and Zotero.

The initial search was conducted in September 2024 in PubMed, Scopus, and CINAHL Complete databases using title and abstract search filters. The search strategy (designed by DS and JO) was broad enough to capture different natural language processing and machine learning methods related to mental health. Table 1 presents the search query for the databases.

Table 1. Search strategy for PubMed and Scopus.

("natural language processing" OR "large language model*" OR "LLM" OR "NLP" OR "ChatGPT" OR "GPT-3" OR "GPT-4" OR "Llama" OR "Mistral" OR "BARD" OR "Mixtral" OR "transformer*" OR "Gemini" OR "Copilot" OR "BERT" OR "RoBERTa" OR "ALBERT" OR "Claude" OR "text mining" OR "text extraction" OR "Generative AI" OR "Natural language understanding" OR "GLoVe" OR "text2vec" OR "doc2vec" OR "word2vec" OR "fastText" OR "attention mechanism" OR "sequence-to-sequence models" OR (("CNN" OR "neural network*" OR "GRU" OR "Gated Recurrent Unit" OR "Long Short-Term Memory" OR "RNN" OR "LSTM" OR "DNN" OR "deep learning" OR "SVM" OR "support vector machine*" OR "gradient boosting" OR "LASSO" OR "XGBoost" OR "AdaBoost" OR "random forest" OR "regression" OR "machine learning") AND "text")) AND ("mental" OR "well-being" OR "depression" OR "anxiety" OR "social work" OR "psychology" OR "psychiatry" OR "abuse" OR "violence" OR "addiction" OR "suicide" OR "grief" OR "bereavement" OR "trauma" OR "stressful life events" OR "counseling") AND ("database" OR "dataset" OR "repository" OR "corpus" OR "collection" OR "reports" OR "discharge summaries" OR "documents" OR "records" OR "patient summaries" OR "notes" OR "text" OR "texts")

All citations were uploaded to Covidence software which was used to track progress of the project in lieu of the protocol [21]. The screening and extraction process took place in Covidence. The specific method we used to conduct this review was automating the process of screening and extraction with the help of the LLM plugin for Covidence that we developed. The process of using

LLM for screening and extraction is depicted in *Figure 1*.

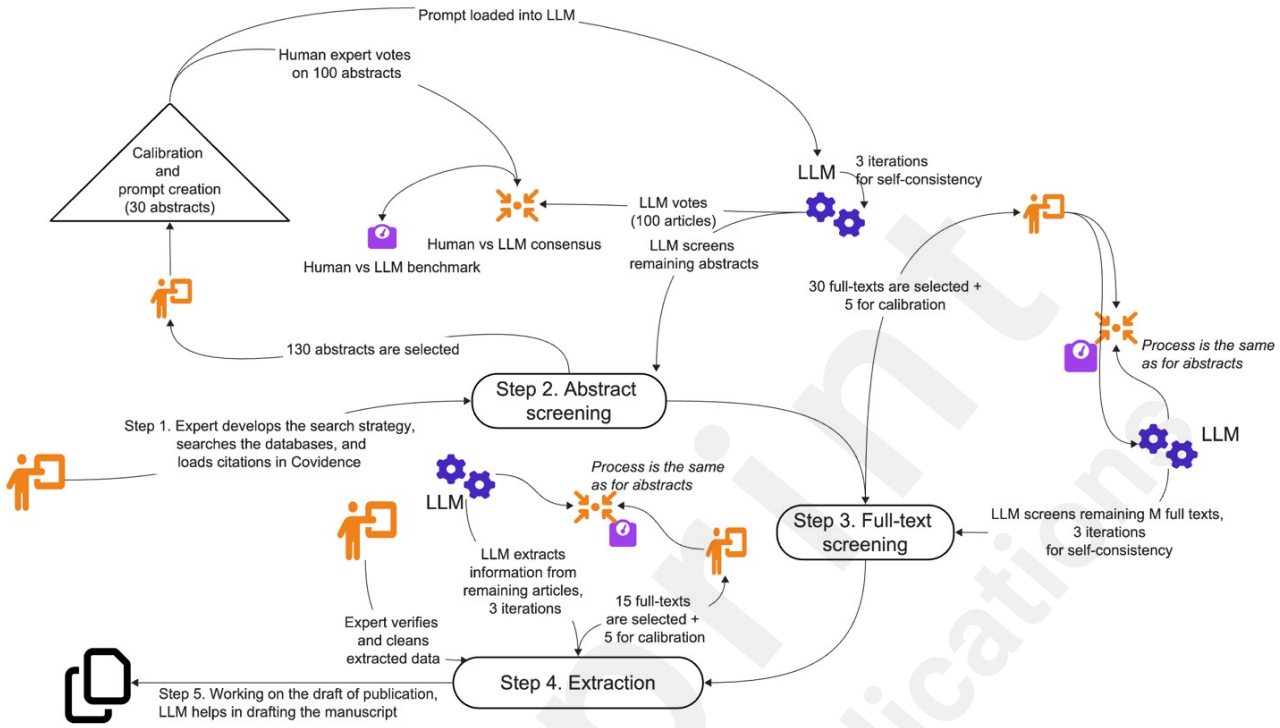

Figure 1. LLM add-on assistant in Covidence screening.

Our add-on for Covidence works by interacting with Covidence using R scripts with the Selenium automation package. Our scripts, which can run both locally or on the server, pass contents between Covidence and Azure OpenAI LLM using OpenAI Application Programming Interface (API) wrapped in the Python “openai” library. Specific models we used in this study are gpt-4o and gpt-4o-mini. Once the model generates the response, our add-on automates actions in Covidence, such as clicking the Include/Exclude buttons. Our automation scripts require a PDF file of the full text to be uploaded into Covidence for full-text screening and extraction. The LLM module supports non-English languages natively. Our software product is still in development and not available to the public, but we list the LLM prompts for each phase of the review in Supplementary Appendix S2. A similar approach was described in our recent paper on the usage of LLM for review automation [22].

The resulting process can be described as follows. Each of the three main stages of review in Covidence (abstract screening, full-text screening, and extraction) included a calibration phase, where a human reviewer (DS) experienced in conducting scoping and systematic reviews screened a small sample of abstracts or full-text articles to get a deeper understanding of inclusion criteria and extraction categories; then we created a prompt for the LLM and tested the LLM performance on

another sample of abstracts or full-texts. Three repeated requests to LLM with the same prompt are made by our automation scripts, and the majority vote principle is used to determine the LLM vote (e.g. LLM votes for an abstract: “include”, “exclude”, “include” means LLM final vote is “include”). Out of these three requests, two are made using gpt-4o-mini and one using more powerful gpt-4o for screening; all three requests are using gpt-4o-mini for extraction to cut API costs.

During the benchmark, a human reviewer (DS) compared his votes against LLM and produced a new gold standard label (human-LLM consensus), and against this consensus, initial human expert and LLM votes were measured. Extraction precision was measured using a simplified benchmark where LLM results on 15 publications were checked by human reviewer (DS) for precision only. The benchmarks are available in Supplementary Appendix S3.

The data charting form for extraction was designed by human experts (DS and JO) and adopted into LLM prompt to collect the following primary information:

- Author, year, title;
- Country or US state (if it is in the US) where the study was conducted or first author's affiliation location
- Natural language processing (NLP) method that was used (generally described in the Methods section), e.g., recurrent neural network (RNN), convolutional neural network (CNN), random forest, deep learning, pattern-matching, ChatGPT, GPT-4, etc.
- What mental health problem(s) were investigated in the paper?
- What is the mental health area or specialty that best represents this paper (psychology, well-being, psychiatry, social work, substance abuse, marriage therapy, addiction therapy, suicide, grief, bereavement, trauma, stressful life events, counseling, other)
- Variables used in the study related to demographics, for example, age, race, ethnicity, gender, sex at birth, marital status, relationship status, sexual orientation, etc.
- Variables used in the study related to social determinants of health, such as none mentioned, urban/rural, transportation availability, access to healthcare, incarceration, income, poverty, health insurance, language knowledge, living arrangement, children/ childless, family, adverse childhood experiences, housing, education, religion, stress, traumatic events, stressful life events, etc.
- Name of the text dataset that was used in the study
- What is the type of this text dataset (e.g., clinical notes, therapy session notes, social media platforms, online forum, other)
- What information or variables were extracted from this text dataset?

- Is it mentioned in the paper if it is possible for other researchers to get access to this text dataset?
- If it is mentioned in the paper that it is possible to get access to this text dataset, what kind of access is it (e.g., public, public with restrictions, private, not given, not mentioned)? If the dataset can be found online or in well-known competition platforms like Kaggle, it is considered public.
- If it is mentioned in the paper that access to this text dataset is public or public with restrictions, what is required to get access (can be training, signing a use agreement, emailing the author, or similar)?
- a Uniform Resource Locator (URL) to the text dataset, if provided. URL was additionally validated using an R script to test if "OK" reply is returned by the server.

Extracted results were synthesized using a table with a complete list of all citations, using maps for location information, and column plots displaying frequency statistics for all other extracted variables.

ChatGPT was used to clean the extraction data, specifically, format the case, remove duplicates, and sort entries into higher-level groups. Scite.ai was used to draft parts of the introduction and discussion sections, while ChatGPT was used to draft the abstract and results section of this review by generating text and R code snippets, which were then corrected (DS) where needed. Human experts (DS and JO) edited and verified LLM-generated parts of the manuscript. Due to the significant number of reviewed citations, publication information, such as authors, title, and DOI, is provided in Supplemental Appendix S4.

Results

Figure 2 shows the flow diagram of the scoping review process. Initially, 12,901 articles were identified from various databases and registers: 6,791 from Scopus, 5,500 from PubMed, and 610 from CINAHL Complete. After the removal of 1,027 duplicates by Covidence, 11,878 studies were retained for screening.

During the abstract and title screening phase, 8,197 studies were excluded, leaving 3,681 studies for retrieval. Out of these, 1,649 studies could not have their full-text PDFs retrieved using automated tools like Covidence and EndNote. Consequently, 2,032 studies were assessed for eligibility. Of these, 264 studies were excluded for the following reasons: 217 were not focused on one of the mental health areas, 39 did not use any kind of natural language processing method, 1 study lacked sufficient information or was too brief, and 2 were review papers (systematic, scoping,

literature, narrative, or other types of reviews), and 5 additional duplicates were identified during full-text screening. The final review included 1,768 studies.

Figure 3 illustrates the geographic distribution of studies reviewed in this analysis. The total number of studies reviewed. Most studies originated from the USA, with 624 studies (35.3%). China contributed 197 studies (11.1%), followed by the United Kingdom (n=167, 9.4%), and India (n=120, 6.8%).

Canada contributed 51 studies (2.9%). Other notable contributors include Japan with 49 studies (2.8%), Spain with 39 studies (2.2%), Australia with 38 studies (2.1%), South Korea (n=27, 1.5%), Germany (n=26, 1.5%), the Netherlands (n=25, 1.4%), Saudi Arabia (n=24, 1.4%), Italy (n=22, 1.2%), and France (n=21, 1.2%), and several other countries each contributing between 1 and 16 studies.

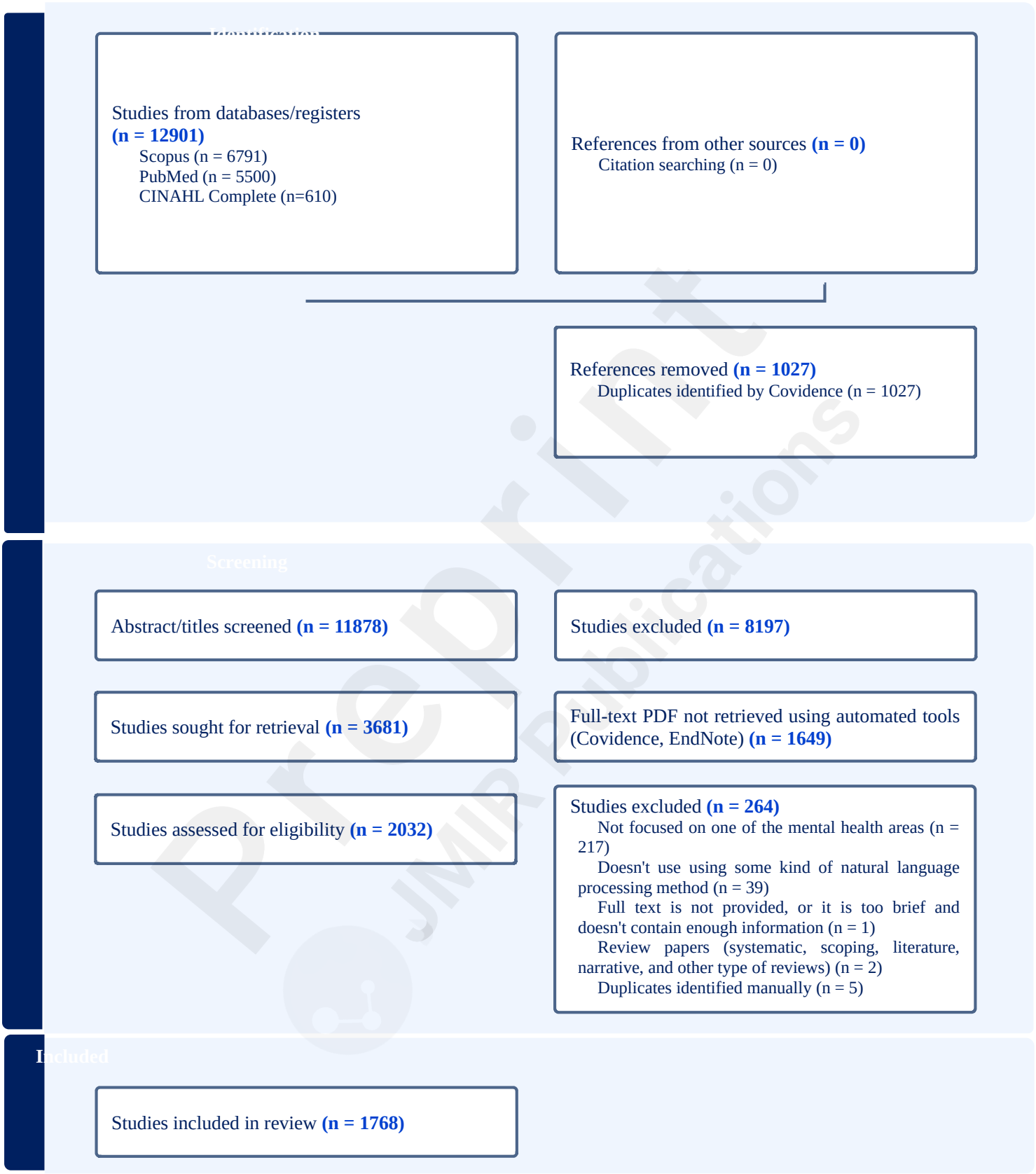

Figure 2. Flow diagram of the scoping review process.

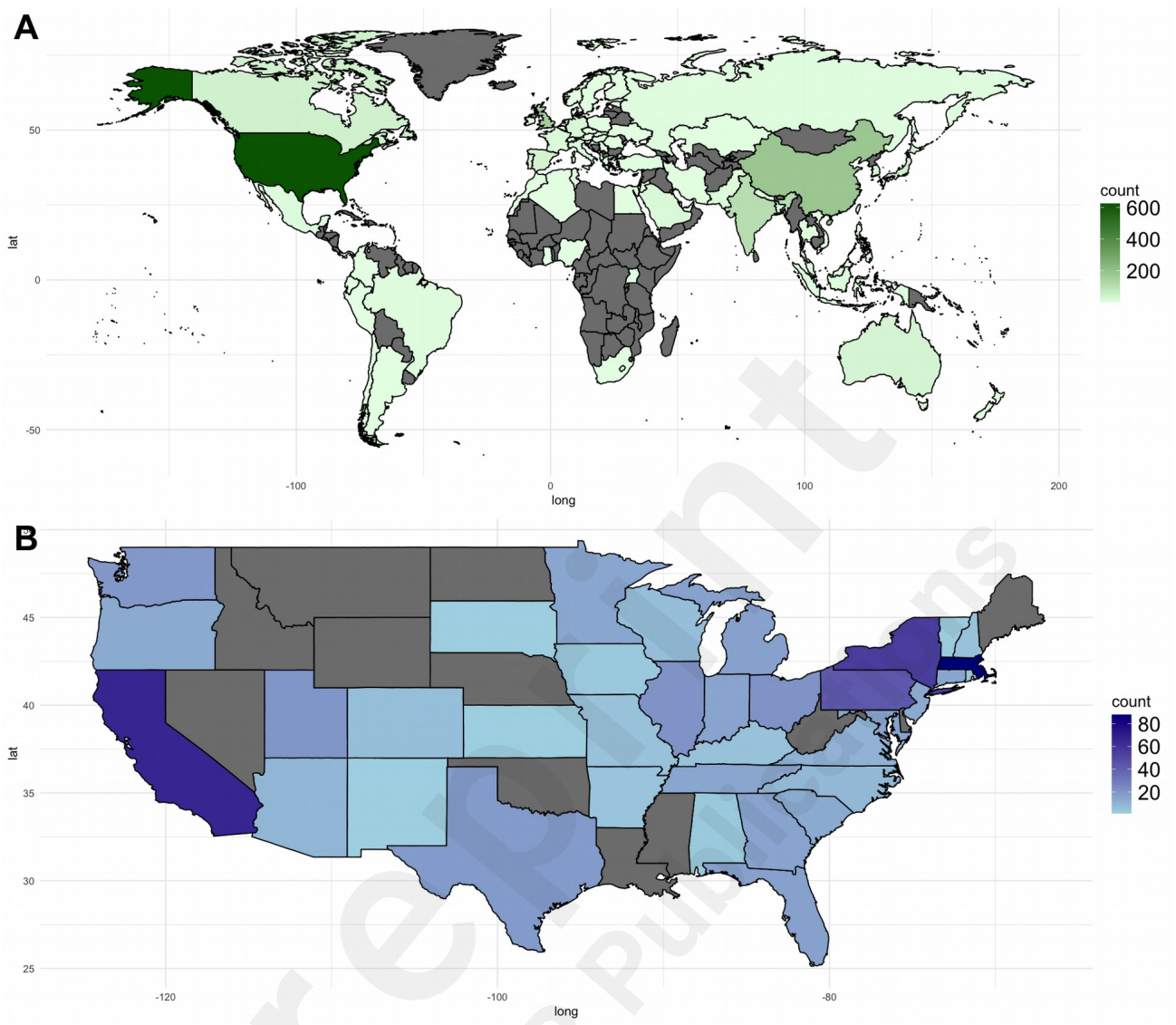

Figure 3. A: Publications by country of origin; B: Publications by state in the US.

Specifically, within the USA, Massachusetts led the contributions with 88 studies (14.1%), followed by California with 66 studies (10.6%). New York contributed 55 studies (8.8%), while Pennsylvania provided 43 studies (6.9%). Ohio and Illinois each contributed 21 studies (3.4%). Other states such as Utah, Washington, and Texas contributed 20 (3.2%), 19 (3.0%), and 18 studies (2.9%), respectively. Michigan and Florida each added 16 studies (2.6%), while Maryland, Georgia, and Indiana each contributed 15 studies (2.4%). Tennessee, Minnesota, and Connecticut each contributed 14 studies (2.2%), followed by New Jersey, Oregon, and South Carolina with 13 studies each (2.1%). Other states contributed fewer than 10 studies.

Figure 4A illustrates the various mental health topics covered in the reviewed papers. The most frequently discussed topic is Depression, which is covered in 518 papers (29.3%). This is

followed by Suicide, featured in 273 papers (15.4%), and Anxiety, discussed in 202 papers (11.4%). Substance Use Disorder is also a significant topic, appearing in 166 papers (9.4%), while Mental Health (Unspecified) is mentioned in 120 papers (6.8%).

Other notable topics include Stress (n=62, 3.5%), Dementia (n=59, 3.3%), Post-Traumatic Stress Disorder (PTSD) (n=53, 3.0%), and Schizophrenia (n=53, 3.0%). Bipolar Disorder appears in 43 papers (2.4%), and Domestic Violence is discussed in 29 papers (1.6%). Eating Disorders are mentioned in 26 papers (1.5%), while Cyberbullying and Cancer-Related topics are covered in 23 (1.3%) and 22 papers (1.2%), respectively. Self-Harm is discussed in 21 papers (1.2%), and Loneliness is covered in 19 papers (1.1%).

Additionally, other mental health issues like Attention-Deficit/Hyperactivity Disorder (ADHD) (n=18, 1.0%), Psychosis (n=18, 1.0%), Autism Spectrum Disorder (n=17, 1.0%), and Diabetes-related mental health issues (n=13, 0.7%) are also represented. Topics such as Pain (n=12, 0.7%) and Fear (n=11, 0.6%) are covered, along with Personality Traits (n=11, 0.6%) and Burnout (n=8, 0.5%). A variety of other mental health topics appear in 1-7 papers.

Figure 4B illustrates the various methodologies and tools discussed in the automation papers. The most frequently mentioned are Neural Network Models (n=499, 28.2%), which include examples such as CNN, LSTM (Long Short-Term Memory), BI-LSTM-CNN (Bidirectional LSTM CNN), GRU (Gated Recurrent Unit), and RNN. Other Machine Learning Models are discussed in 355 papers (20.1%), highlighting the use of Random Forest, Support Vector Machine (SVM), regression models, and Gradient Boosting Trees.

Transformer Models appear in 312 papers (17.6%), with examples like BERT, GPT-3, LLAMA-2, and Roberta. Natural Language Processing (NLP) Tools are featured in 264 papers (14.9%), utilizing tools such as Spacy NLP Library, Stanford CoreNLP, and GATE for processing and analyzing text data. Topic Modeling & Text Mining are discussed in 258 papers (14.6%), employing techniques such as Latent Dirichlet Allocation (LDA), Structural Topic Modeling (STM), and Biterm Topic Modeling (BTM) for extracting themes and patterns from text data.

Traditional Text Representation & Embedding methods are mentioned in 90 papers (5.1%), including methods like Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and N-gram Representation. Unspecified Machine Learning approaches appear in 61 papers (3.4%), while Sentiment Analysis is discussed in 31 papers (1.8%). Lastly, Linguistic Inquiry & Word Count

(LIWC) is mentioned in 22 papers (1.2%), showcasing tools such as LIWC15 Text Analysis and LIWC Dictionaries. Rule-Based Methods are included in 15 papers (0.8%).

Sentiment Analysis is discussed in 30 papers (1.9%), with approaches like VADER, Aspect-Based Sentiment Analysis, and Text Sentiment Analysis. Rule-Based Methods are featured in 15 papers (1.0%), using approaches such as Pattern-Matching and Lexicon-Based NLP to perform specific text processing tasks based on predefined rules. Finally, Bayesian Models are mentioned in 3 papers (0.2%), where techniques like Bayesian Networks and Bayesian Logistic Regression are applied, indicating a more niche focus on this approach within the reviewed literature. The Other category, covered in 289 papers (18.6%), represents a wide range of techniques beyond the most common methods, including various specialized or less frequently used approaches.

Figure 4C presents an overview of the types of datasets utilized in the reviewed studies. The most commonly used dataset type is Clinical Data, which appears in 751 papers (42.4%), followed by Social Media datasets with 592 papers (33.4%). Online Forums have substantial representation as well, with 89 papers (5.0%), and the Other category comprises 99 papers (5.6%). Survey Data is also notable, appearing in 23 papers (1.3%), while Mobile and Digital Health Data is used in 21 papers (1.2%).

Less frequently used datasets include Counseling Data (n=14, 0.8%), Audio and Video Data (n=14, 0.8%), and Chatbot and AI Interaction Data (n=8, 0.5%). The Articles and Academic Texts category is represented in 9 papers (0.5%), while Websites and Online Platforms account for 7 papers (0.4%). Blogs and Online Articles have 4 papers (0.2%).

Other datasets like Diary and Personal Account Data and Synthetic Data each appear in 2 papers (0.1%), along with Focus Groups, which are represented in 3 papers (0.2%). Lastly, YouTube Data is noted in 1 paper (<0.1%), indicating niche areas of study or emerging methodologies within the broader field.

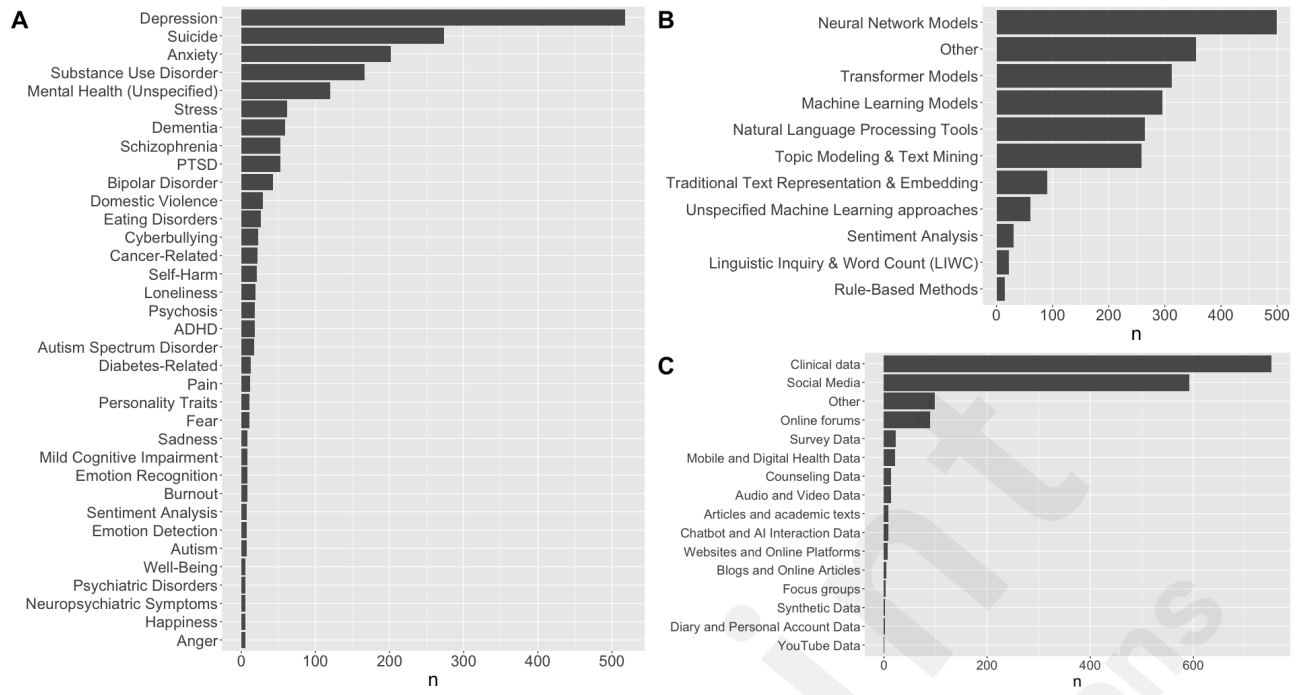

Figure 4 A: Mental health outcomes studied in reviewed publications (mentioned in >= 5 citations), *B:* NLP methods/tools used (mentioned in >= 5 citations), *C*: Types of datasets used for analysis

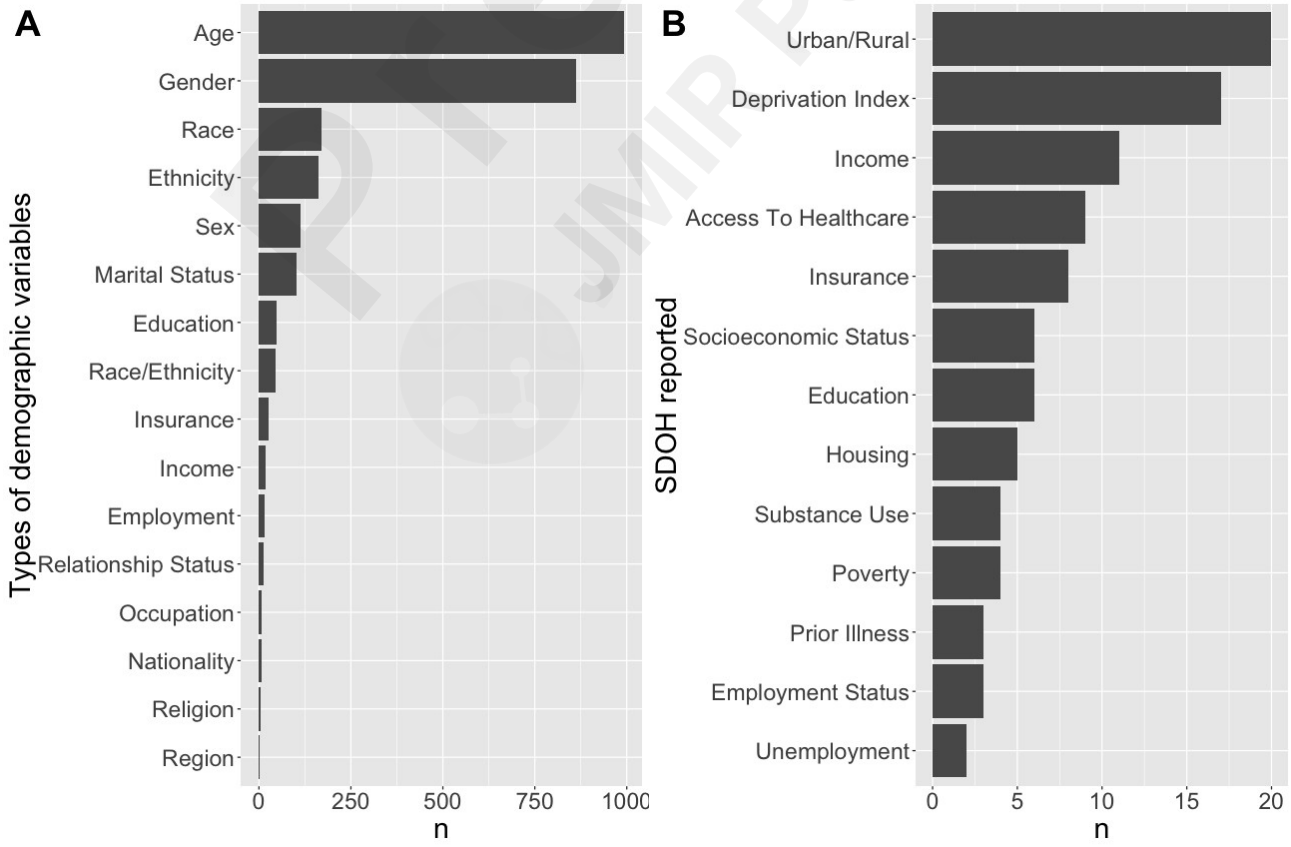

Figure 5 A: Demographic variables used in the studies (with >=3 mentions), *B:* Social determinants used in the studies (>= 2 mentions).

Figure 5A presents a detailed overview of the demographic variables frequently utilized in the reviewed studies. Age is the most commonly extracted variable, appearing in 993 studies (56.2%). Following closely is Gender, featured in 863 studies (48.8%). Other significant demographic variables include Race (n=171, 9.7%) and Ethnicity (n=161, 9.1%). Sex is documented in 114 studies (6.4%), while Marital Status appears in 101 studies (5.7%).

Less frequently reported variables include Education (n=48, 2.7%), Race/Ethnicity (n=46, 2.6%), and Insurance (n=26, 1.5%). Income is mentioned in 19 studies (1.1%), with Employment in 15 studies (0.8%), Relationship Status in 13 studies (0.7%), and Occupation in 8 studies (0.5%). More specialized demographic insights are provided by variables like Nationality (n=6, 0.3%), Religion (n=4, 0.2%), and Region (n=3, 0.2%). Additionally, niche variables such as Aboriginal Status, Career, and Socioeconomic Status are noted, each appearing in 2 studies (0.1%).

It is important to highlight that demographic variables experienced a notable number of false positives during extraction, with a precision rate of 0.66, suggesting that the actual counts for gender and age may be significantly lower.

Figure 5B offers an overview of the social determinants of health variables utilized in the reviewed studies. The Urban/Rural status is the most frequently reported variable, appearing in 20 studies (1.2%). Following closely is the Deprivation Index, included in 17 studies (1.1%). Income is mentioned in 11 studies (0.7%), underscoring its significance in assessing economic conditions.

The relevance of Access to Healthcare and Insurance is reflected in their occurrence in 9 studies (0.5%) and 8 studies (0.5%), respectively. Education and Socioeconomic Status are recorded in 6 studies (0.4%) each, while Housing is featured in 5 studies (0.3%).

Less frequently reported variables include Poverty and Substance Use, each appearing in 4 studies (0.3%), as well as Employment Status and Prior Illness, each in 3 studies (0.2%). Additional variables such as Unemployment (n=2, 0.1%) and various specific factors—like Domestic Violence, Drug Involvement, and others—are noted in just 1 study (0.06%) each.

Figure 6 illustrates the most frequently extracted information from the text datasets. The scope of extracted data includes information related to sentiments and emotions, health conditions, health symptoms, personality traits, violence and bullying, suicide indicators, user engagement

(likes, shares), survey data, and language features, to name a few.

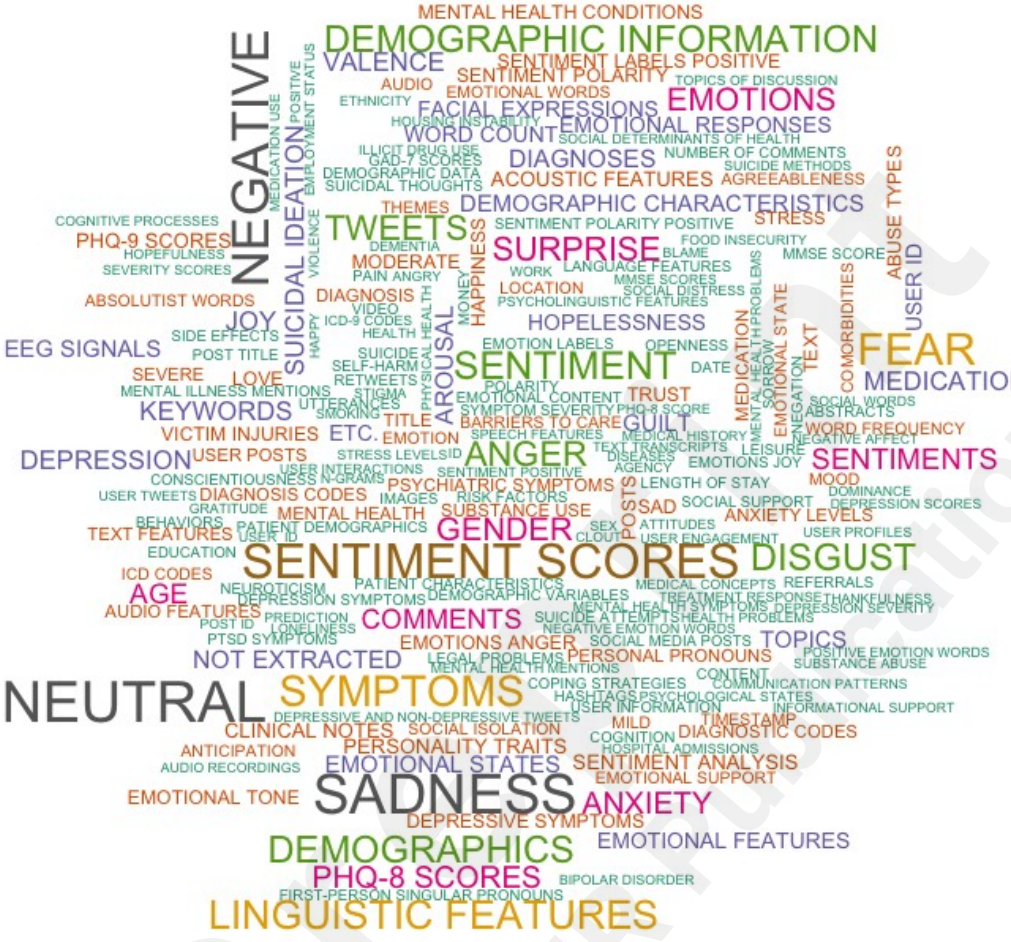

Figure 6 Word cloud showing a range of information extracted from text datasets (terms with >= 3 mentions are displayed)

Figure 7 illustrates a notable disparity in the availability of datasets across the studies reviewed. A significant majority of studies, 911 (51.5%), did not clarify whether their datasets were accessible, while 857 studies (48.5%) included access information for their datasets.

Regarding the specific levels of access, the vast majority of studies fell into the “Unclear” category, with 1,128 studies (63.7%) failing to provide explicit information about dataset accessibility. In contrast, 362 studies (20.5%) indicated that their datasets were publicly accessible,

while 263 studies (14.9%) allowed public access with certain restrictions, thus enabling data use under specific conditions. Only a minimal number of studies categorized their datasets as private, with just 9 studies (0.5%) restricting access to particular individuals or groups. Additionally, 4 studies (0.2%) did not provide any information regarding their dataset access levels.

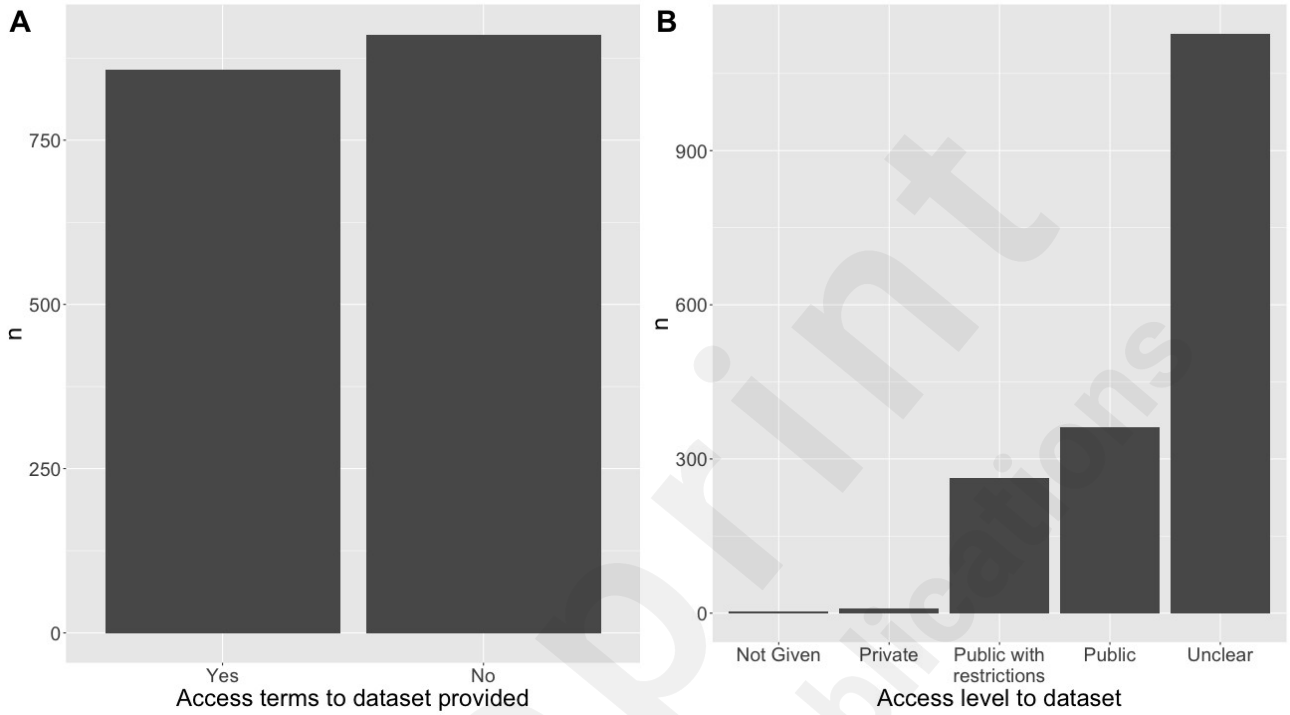

Figure 7 Citations by *A*: access terms to dataset provided in the paper, *B:* access level to dataset mentioned

Additional information on the text datasets, access levels, and other extracted information can be found in Supplemental Appendix S4.

Discussion

Our LLM-assisted scoping review revealed a wide range of projects related to mental health outcomes that use NLP. The US leads the race in this domain, with more than a third of the publications appearing in this country, but China, the UK, and India follow closely behind, reflecting the worldwide interest in the application of NLP to mental health problems.

Depression emerged as the top mental health problem investigated, reflecting the current trends: a study by Wang et al. [23] utilizing Google Trends data in the United States reported a 61% increase in depression prevalence from 2008 to 2018. Another study by Jonson et al. [24] in Sweden observed a decline in depression prevalence among 85-year-olds, potentially influenced by rising trends in younger age groups. Suicide was especially well represented in our sample, highlighting the

fact that suicide mortality remains a significant global public health concern. This finding is echoed by studies indicating that suicide continues to be a notable contributor to mortality worldwide [25].

Neural networks appear to dominate the landscape of tools used in NLP research, highlighting their versatility and performance. They are known for their ability to effectively learn from samples, mimic human brain functions, approximate nonlinear mappings, and establish complex system models [26]. The self-organizing, self-learning, and parallel distributed information processing capabilities of neural networks have made them invaluable in pattern recognition, signal processing, and optimization problems [27]. Moreover, artificial neural networks are recognized for their versatility in solving nonlinear problems with multiple independent variables [28].

Clinical data and social media dominate the types of datasets used in the projects, showing two major avenues of NLP mental health research, one with medical records data and the other with using public social media platforms.

As for SDOH and demographic variables, there is considerable overlap between the two in the extracted data. Previous work suggests that demographic variables should be part of SDOH, for example, the commonly used variable marital status reflects the social connections, stage of life, and other important social implications for individuals' health [29]. The same can be said about age, the most frequently reported demographic variable. Research has shown that disparities in mental health outcomes persist across different age groups and are often linked to social stress, discrimination, and stigma [30]. These disparities can be exacerbated by obstacles to healthcare access based on factors like ethnicity, sex, and occupation [31].

Perhaps we should rethink the distinction between the demographic variables and SDOH. This study suggests that social determinants, besides gender and age, were rarely used in the studies, highlighting a significant gap that could be addressed in future work. In addition, a manual review of LLM outputs (see Supplementary Appendix S3 for benchmark) revealed that the demographic variables category had a lot of false positives, which means that gender and age were actually used even less frequently in the studies, most commonly they were reported in the introduction sections as important factors and ignored in the actual analysis.

Our method proved to be sensitive at detecting relevant citations and fetched our own previous work on suicide, self-harm, opioid addiction, and other topics [32-40]. Future work could be facilitated by this review, which revealed a considerable number of research datasets and provided URLs (for some of them). In fact, over 600 publications disclosed the datasets they used with the level of access mentioned as public and semi-public. Since finding new data for research is always challenging, we hope that this review can serve as a starting point in mental health NLP research.

Limitations

We used calibrated LLMs as assistants in this review project. Some extraction categories, such as performance metrics, had relatively lower accuracy, so the results of this extraction category should be taken with caution. Nevertheless, in this review, LLM achieved remarkable results in accuracy, making it possible to delegate time-consuming phases of review to LLM while leaving humans the supervision and benchmark of the process. LLM also supports non-English languages natively, which allowed us to capture more diverse range of works.

This study only utilized single human reviewer assisted by LLM. Studies generally recommend a single reviewer approach in some cases, like rapid reviews [41], however, we believe that the LLM approach could substitute human reviewers, and human effort should be redirected to the supervision of the review process.

Our method relies on the PDF of the publication, for a considerable number of papers, we could not locate a PDF using citation manager tools, such as EndNote and Zotero. Thus, we had to exclude a considerable number of papers from our analysis. However, we believe that the number of full-text citations that we obtained was large enough to get a statistical representation of the extracted categories and to support our position elaborated in the discussion section.

Conclusions

This review highlights the range of projects using NLP for mental health areas, with depression and suicide being the most frequent health outcomes under study. Social determinants were rarely used in the publications, with traditional demographic variables, such as age and gender, being more frequent. The extracted information could be used by other researchers looking for text datasets or for projects in specific areas.

Author contributions

All authors conceived and designed the study. DS and JO contributed to search strategy development. DS performed the benchmarks for LLM, prompt engineering, and verified data extraction results. All authors contributed to the interpretation and drafted the manuscript. All authors critically reviewed and revised the manuscript and approved the final version for submission.

Funding

This publication was supported, in part, by the National Center for Advancing Translational Sciences of the National Institutes of Health under Grant Number UL1 TR001450. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This publication was supported in part by a Smart-state Chair

endowment.

Data availability

The data underlying this article are available in the article and its online supplementary material. The code used for data cleaning is available on GitHub [42].

Conflicts of interest statement

The authors have no competing interests to declare.

References

1. Sedgwick R, Bittar A, Kalsi H, Barack T, Downs J, Dutta R. Investigating Online Activity in UK Adolescent Mental Health Patients: A Feasibility Study Using a Natural Language Processing Approach for Electronic Health Records. BMJ Open. 2023;13(5):e061640.
2. Ridgway JP, Uvin AZ, Schmitt J, Oliwa T, Almirol E, Devlin S, et al. Natural Language Processing of Clinical Notes to Identify Mental Illness and Substance Use Among People Living With HIV: Retrospective Cohort Study. Jmir Medical Informatics. 2021;9(3):e23456.
3. Houssein EH, Mohamed RE, Ali AA. Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review. Ieee Access. 2021;9:140628-53.
4. Zirikly A, Desmet B, Newman-Griffis D, Marfeo E, McDonough CM, Goldman HH, et al. Information Extraction Framework for Disability Determination Using a Mental Functioning Use-Case. Jmir Medical Informatics. 2022;10(3):e32245.
5. Calvo RA, Milne D, Hussain MS, Christensen H. Natural Language Processing in Mental Health Applications Using Non-Clinical Texts. Natural Language Engineering. 2017;23(5):649-85.
6. Glaz AL, Haralambous Y, Kim-Dufor D-H, Lenca P, Billot R, Ryan TC, et al. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. Journal of Medical Internet Research. 2021;23(5):e15708.
7. Bieri JS. Natural Language Processing for Work-Related Stress Detection Among Health Professionals: Protocol for a Scoping Review. Jmir Research Protocols. 2024;13:e56267.
8. Chandran D, Robbins DA, Chang CK, Shetty H, Sanyal J, Downs J, et al. Use of Natural Language Processing to Identify Obsessive Compulsive Symptoms in Patients With Schizophrenia, Schizoaffective Disorder or Bipolar Disorder. Scientific Reports. 2019;9(1).
9. Campion ED. Using Natural Language Processing to Increase Prediction and Reduce Subgroup Differences in Personnel Selection Decisions. Journal of Applied Psychology. 2024;109(3):307-38.
10. Deferio JJ, Breitinger S, Khullar D, Sheth AP, Pathak J. Social Determinants of Health in Mental Health Care and Research: A Case for Greater Inclusion. Journal of the American Medical Informatics Association. 2019;26(8-9):895-9.
11. Gutiérrez D, Conley AH. What Just Is Isn't Always Justice: Toward a Spiritual View of Justice. Counseling and Values. 2021;66(2):114-6.
12. Dolatabadi E, Moyano D, Bales ME, Spasojević S, Bhambhoria R, Bhatti JA, et al. Using Social Media to Help Understand Patient-Reported Health Outcomes of Post–COVID-19 Condition: Natural Language Processing Approach. Journal of Medical Internet Research. 2023;25:e45767.
13. Dorr DA, Quiñones A, King T, Wei MY, White K, Bejan CA. Prediction of Future Health Care Utilization Through Note-Extracted Psychosocial Factors. Medical Care. 2022;60(8):570-8.

14. Raza S, Schwartz B. Constructing a Disease Database and Using Natural Language Processing to Capture and Standardize Free Text Clinical Information. Scientific Reports. 2023;13(1).
15. Rouillard C, Nasser MA, Hu H, Roblin DW. Evaluation of a Natural Language Processing Approach to Identify Social Determinants of Health in Electronic Health Records in a Diverse Community Cohort. Medical Care. 2022;60(3):248-55.
16. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. International journal of social research methodology. 2005;8(1):19-32.
17. National Institute for Health and Care Excellence. Use of AI in evidence generation: NICE position statement.
18. Peters MD, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI evidence implementation. 2021;19(1):3-10.
19. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of internal medicine. 2018;169(7):467-73.
20. Lindley L, Galupo MP. Gender dysphoria and minority stress: Support for inclusion of gender dysphoria as a proximal stressor. Psychology of Sexual Orientation and Gender Diversity. 2020;7(3):265.
21. Covidence. Available from: https://www.covidence.org/.
22. Scherbakov D, Hubig N, Jansari V, Bakumenko A, Lenert LA. The emergence of Large Language Models (LLM) as a tool in literature reviews: an LLM automated systematic review. arXiv preprint arXiv:240904600. 2024.
23. Wang A, McCarron RM, Azzam D, Stehli A, Xiong GL, DeMartini J. Utilizing Big Data From Google Trends to Map Population Depression in the United States: Exploratory Infodemiology Study. Jmir Mental Health. 2022;9(3):e35253.
24. Jonson M, Sigström R, Hedna K, Sterner TR, Erhag HF, Wetterberg H, et al. Time Trends in Depression Prevalence Among Swedish 85-Year-Olds: Repeated Cross-Sectional Population-Based Studies in 1986, 2008, and 2015. Psychological Medicine. 2021;53(6):2456-65.
25. Orpana H, Marczak LB, Arora M, Abbasi N, Abdulkader RS, Zegeye A, et al. Global, Regional, and National Burden of Suicide Mortality 1990 to 2016: Systematic Analysis for the Global Burden of Disease Study 2016. BMJ. 2019:l94.
26. Zhao N, Lu J. Review of Neural Network Algorithm and Its Application in Temperature Control of Distillation Tower. Journal of Engineering Research and Reports. 2021:50-61.
27. Pan X, Wang Y, Qian Y. Artificial Neural Network Model and Its Application in Signal Processing. Asian Journal of Advanced Research and Reports. 2023:1-8.
28. Flood I, Bewick BT, Rauch E. Rapid Simulation of Blast Wave Propagation in Built Environments Using Coarse-Grain Based Intelligent Modeling Methods. 2011.
29. Wang Y, Chen Z, Zhou C. Social engagement and physical frailty in later life: does marital status matter? BMC geriatrics. 2021;21:1-11.
30. Bränström R, Fellman D, Pachankis JE. Age-Varying Sexual Orientation Disparities in Mental Health, Treatment Utilization, and Social Stress: A Population-Based Study. Psychology of Sexual Orientation and Gender Diversity. 2023;10(4):686-98.
31. Park D-H, Meltendorf T, Kahl KG, Kamp JC, Richter MJ, Hoeper MM, et al. Health Disparities and Differences in Health-Care-Utilization in Patients With Pulmonary Arterial Hypertension. Frontiers in Psychiatry. 2022;13.
32. Hanson RF, Zhu V, Are F, Espeleta H, Wallis E, Heider P, et al. Initial development of tools to identify child abuse and neglect in pediatric primary care. BMC Medical Informatics and Decision Making. 2023;23(1).
33. Lenert LA, Zhu V, Jennings L, McCauley JL, Obeid JS, Ward R, et al. Enhancing research

data infrastructure to address the opioid epidemic: the Opioid Overdose Network (O2-Net). JAMIA Open. 2022;5(2):ooac055.
34. Obeid JS, Dahne J, Christensen S, Howard S, Crawford T, Frey LJ, et al. Identifying and Predicting Intentional Self-Harm in Electronic Health Record Clinical Notes: Deep Learning Approach. JMIR Med Inform. 2020;8(7):e17784.
35. Obeid JS, Heider PM, Weeda ER, Matuskowitz AJ, Carr CM, Gagnon K, et al. Impact of De-Identification on Clinical Text Classification Using Traditional and Deep Learning Classifiers. Stud Health Technol Inform. 2019;264:283-7.
36. Obeid JS, Tsalatsanis A, Chaphalkar C, Robinson S, Klein S, Cool S, et al. A Reproducible Model Based on Clinical Text for Predicting Suicidal Behavior. Stud Health Technol Inform. 2024;310:1486-7.
37. Obeid JS, Weeda ER, Matuskowitz AJ, Gagnon K, Crawford T, Carr CM, et al. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. BMC Med Inform Decis Mak. 2019;19(1):164.
38. Zhu V, Lenert L, Bunnell B, Obeid J, Jefferson M, Halbert CH. Automatically Identifying Financial Stress Information from Clinical Notes for Patients with Prostate Cancer. Cancer Res Rep. 2020;1(1).
39. Zhu VJ, Lenert LA, Barth KS, Simpson KN, Li H, Kopscik M, et al. Automatically identifying opioid use disorder in non-cancer patients on chronic opioid therapy. Health Informatics Journal. 2022;28(2).
40. Zhu VJ, Lenert LA, Bunnell BE, Obeid JS, Jefferson M, Halbert CH. Automatically identifying social isolation from clinical narratives for patients with prostate Cancer. BMC Med Inform Decis Mak. 2019;19(1):43.
41. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC medical research methodology. 2019;19:1-9.
42. Scherbakov D. Automated-Review-NLP-Datasets-For-Mental-health. 2024 [2024-10-01]. Available from: https://github.com/scherbakovdmitri/Automated-Review-NLP-Datasets-For-Mental-health/tree/main.

Supplementary material

Supplementary Appendix S1. Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
TITLE			
Title	1	Identify the report as a scoping review.	1
ABSTRACT			
Structured summary	2	Provide a structured summary that includes (as applicable): background, objectives, eligibility criteria, sources of evidence, charting methods, results, and conclusions that relate to the review questions and objectives.	1-2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known. Explain why the review	2-4

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
		questions/objectives lend themselves to a scoping review approach.	
Objectives	4	Provide an explicit statement of the questions and objectives being addressed with reference to their key elements (e.g., population or participants, concepts, and context) or other relevant key elements used to conceptualize the review questions and/or objectives.	3-4
METHODS			
Protocol and registration	5	Indicate whether a review protocol exists; state if and where it can be accessed (e.g., a Web address); and if available, provide registration information, including the registration number.	5
Eligibility criteria	6	Specify characteristics of the sources of evidence used as eligibility criteria (e.g., years considered, language, and publication status), and provide a rationale.	4
Information sources*	7	Describe all information sources in the search (e.g., databases with dates of coverage and contact with authors to identify additional sources), as well as the date the most recent search was executed.	4
Search	8	Present the full electronic search strategy for at least 1 database, including any limits used, such that it could be repeated.	5
Selection of sources of evidence†	9	State the process for selecting sources of evidence (i.e., screening and eligibility) included in the scoping review.	6-7
Data charting process‡	10	Describe the methods of charting data from the included sources of evidence (e.g., calibrated forms or forms that have been tested by the team before their use, and whether data charting was done independently or in duplicate) and any processes for obtaining and confirming data from investigators.	7
Data items	11	List and define all variables for which data were sought and any assumptions and simplifications made.	7-8
Critical appraisal of individual sources of evidence§	12	If done, provide a rationale for conducting a critical appraisal of included sources of evidence; describe the methods used and how this information was used in any data synthesis (if appropriate).	Not performed
Synthesis of results	13	Describe the methods of handling and summarizing the data that were charted.	8
RESULTS			
Selection of sources of evidence	14	Give numbers of sources of evidence screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally using a flow diagram.	8-10
Characteristics of sources of evidence	15	For each source of evidence, present characteristics for which data were charted and provide the citations.	9-11, Supplemental Appendix S4
Critical appraisal within sources of evidence	16	If done, present data on critical appraisal of included sources of evidence (see item 12).	Not performed
Results of individual sources of evidence	17	For each included source of evidence, present the relevant data that were charted that relate to the review questions and objectives.	Supplemental Appendix S4
Synthesis of	18	Summarize and/or present the charting results as	11-18

SECTION	ITEM	PRISMA-ScR CHECKLIST ITEM	REPORTED ON PAGE #
results		they relate to the review questions and objectives.	
DISCUSSION			
Summary of evidence	19	Summarize the main results (including an overview of concepts, themes, and types of evidence available), link to the review questions and objectives, and consider the relevance to key groups.	18-20
Limitations	20	Discuss the limitations of the scoping review process.	20
Conclusions	21	Provide a general interpretation of the results with respect to the review questions and objectives, as well as potential implications and/or next steps.	20
FUNDING			
Funding	22	Describe sources of funding for the included sources of evidence, as well as sources of funding for the scoping review. Describe the role of the funders of the scoping review.	21

From: Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. doi: 10.7326/M18-0850.

JBI = Joanna Briggs Institute; PRISMA-ScR = Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews.

* Where *sources of evidence* (see second footnote) are compiled from, such as bibliographic databases, social media platforms, and Web sites.

† A more inclusive/heterogeneous term used to account for the different types of evidence or data sources (e.g., quantitative and/or qualitative research, expert opinion, and policy documents) that may be eligible in a scoping review as opposed to only studies. This is not to be confused with *information sources* (see first footnote).

‡ The frameworks by Arksey and O'Malley (6) and Levac and colleagues (7) and the JBI guidance (4, 5) refer to the process of data extraction in a scoping review as data charting.

§ The process of systematically examining research evidence to assess its validity, results, and relevance before using it to inform a decision. This term is used for items 12 and 19 instead of "risk of bias" (which is more applicable to systematic reviews of interventions) to include and acknowledge the various sources of evidence that may be used in a scoping review (e.g., quantitative and/ or qualitative research, expert opinion, and policy document).

Supplementary Appendix S2. LLM Prompts used for screening and extraction.

Phase of the review	*LLM prompt*

Abstract screening	Summarize the text abstract of a full research paper (article), and given the below criteria list, say if the full paper is likely to be included, excluded, or unclear. Criteria list. Include: Paper should be using some kind of natural processing method (NLP), like transformers, pattern-matching, ChatGPT, GPT-3, BERT, Llama, Mistral, large language models, LDA/LSA, deep learning or machine learning applied to text, and similar. Include: Paper should be in one of the mental health areas, such as: psychology, well-being, psychiatry, social work, substance abuse, marriage therapy, addiction therapy, suicide, grief, bereavement, trauma, stressful life events, counseling, or related. Cyberbullying and study of emotions should be included, however, aggressive and violent language should be excluded. Exclude: If any of the Include criteria doesn't match Exclude: Review papers (systematic, scoping, literature, narrative, and other type of reviews, but retrospective data reviews and chart reviews should be included), book chapters. Exclude: Abstract is not provided, or it is too brief and doesn't contain enough information Follow this format: 1) First provide some explanations why each study should be included or excluded. 2) Then format your output as follows, strictly follow this format, use equal(=) sign, if study is excluded, write 'answer=excluded', if study is included output 'answer=included', or if it is unclear write 'answer=unclear'.

Full-text screening	Look at the research paper (article), and given the below criteria list, say if the full paper is to be included, excluded, or unclear. \nCriteria list. Exclude reason 1: Doesn't use using some kind of natural language processing method (NLP), like transformers, pattern-matching, ChatGPT, GPT-3, BERT, Llama, Mistral, large language models (LLM), LDA/LSA, deep learning or machine learning applied to text, and similar. Exclude reason 2: Not focused on one of the mental health areas, such as: psychology, well-being, psychiatry, social work, substance abuse, marriage therapy, addiction therapy, suicide, grief, bereavement, trauma, stressful life events, counseling, Exclude reason 3: Review papers (systematic, scoping, literature, narrative, and other type of reviews, but retrospective data reviews and chart reviews should be included), conference papers, book chapters Exclude reason 4: Not related to human health or well-being Exclude reason 5: Full text is not provided, or it is too brief and doesn't contain enough information \nFollow this format: 1) First provide some explanations why each study should be included or excluded. 2) Provide citation from text showing what NLP method was used and mental health problem explored. 3) Output the following, choose one best matching exclusion reason: include=yes/no/unclear exclude_reason=reason_number
Extraction of data	Definition of text dataset: Text dataset (also called corpus, notes collection, notes database, text archive, text compilation, text repository, or similar term can be used) is a collection of texts or notes that were used for natural language processing (NLP) analysis or for training NLP models, or for data extraction (also called text mining). Look at the research paper (article), and extract the following information. Follow the format given.\n Field 1) Extract country and US state (if it is in US) where study location was. Typically it is the location where dataset comes from as described in the methods section. If this can not be determined, look at the country and US state of first author's affiliation. Output as: Country name, or USA/State Name Field 2) What natural language processing (NLP) method was used (generally described in Methods section), example answers: the study didn't use natural language processing, word2vec, text2vec, doc2vec, RNN, CNN, SVM, random forest, deep learning, pattern-matching, ChatGPT, GPT-4, BERT, Llama, Mistral, LDA/LSA, other (provide name). Field 3) What mental health problem(s) were investigated in the paper? Field 4) What is the mental health area or specialty that best represents this paper, select one of: not related to mental health, psychology, well-being, psychiatry, social work, substance abuse, marriage therapy, addiction therapy, suicide, grief, bereavement, trauma, stressful life events, counseling, other (provide name). Field 5) List all variables used in the study related to demographics, for example: age, race, ethnicity, gender, sex at birth, marital status, relationship status, sexual orientation, etc. Field 6) List all variables used in the study related to social determinants of health, such as: none mentioned, urban/rural, transportation availability, access to healthcare, incarceration, income, poverty, health insurance, language knowledge, living arrangement, children/childless, family, adverse childhood experiences, housing, education, religion, stress, traumatic events, stressful life events, etc. The next fields are all related to the text dataset that was used in the study: Field 7) What is the name of the text dataset that was used for the Methods section (not to be confused with Introduction)

	Field 8) What is the type of this text dataset, select one of: clinical notes, therapy session notes, social media platforms, online forum, other [insert type here]? Field 9) What information or variables were extracted from this text dataset? Field 10) Is it mentioned in the paper if it is possible for other researchers to get access to this text dataset? Field 11) If it is mentioned in the paper that it is possible to get access to this text dataset, what kind of access it is? Select one of: public, public with restrictions, private, not given, not mentioned Terms of access to text dataset can sometimes be found in the methods section, sometimes in data availability section, however this section has to specifically mention the text dataset that was used in this study. Sometimes terms of access are found in other parts of the document. If the dataset can be found online or in well-known competition platforms like Kaggle consider access as public. Field 12) If it is mentioned in the paper that access to this text dataset is public or public with restrictions, what is required to get access (can be training, signing use agreement, emailing the author, or similar)? Field 13) Link (URL) to the text dataset, if provided. Format your output as an R data.frame: data.frame(fld1='',fld2='',fld3='',...,fld13='')

Supplementary Appendix S3. LLM benchmarks

Table 1. Benchmark of abstract screening phase (N=100 abstracts).

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
Human reviewer vs Consensus	0.89	0.91	0.92	0.87	0.92	0.89	0.91	0.54	0.48	0.52	0.9
LLM vs Consensus	0.98	0.96	0.96	0.98	0.96	0.98	0.97	0.54	0.53	0.55	0.97

Table 2. Benchmark of full-text screening phase (N=30 full-text PDFs).

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
Human reviewer vs Consensus	0.7	1	1	0.87	1	0.7	0.82	0.33	0.23	0.23	0.85
LLM vs Consensus	0.7	0.95	0.87	0.86	0.88	0.7	0.78	0.33	0.23	0.27	0.82

Table 3. Benchmark of full-text extraction phase (N=15 full-text PDFs).

	Country	NLP method	Mental health Outcome	Demographic variables	SDOH Variables	Dataset name	Dataset type	Citation to support authors opinion	Is access to dataset discussed?	Access level mentioned	Requirements to access dataset	URL to dataset	Average
Precision	0.93	0.86	0.93	0.66	1	0.93	0.94	0.93	0.8	0.73	0.93	0.73	*0.86*

Supplementary Appendix S4. The complete extraction table is available as an Excel file. Asterisks near terms denote LLM certainty (*** match across 3 runs, ** match across 2 runs , * term matched only in one LLM run).

Supplementary Files