npj | digital medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Check for updates

Natural language processing applied to mental illness detection: a narrative review

Tianlin Zhang[1], Annika M. Schoene[1], Shaoxiong Ji[2] and Sophia Ananiadou[1,3]✉

Mental illness is highly prevalent nowadays, constituting a major cause of distress in people's life with impact on society's health and well-being. Mental illness is a complex multi-factorial disease associated with individual risk factors and a variety of socioeconomic, clinical associations. In order to capture these complex associations expressed in a wide variety of textual data, including social media posts, interviews, and clinical notes, natural language processing (NLP) methods demonstrate promising improvements to empower proactive mental healthcare and assist early diagnosis. We provide a narrative review of mental illness detection using NLP in the past decade, to understand methods, trends, challenges and future directions. A total of 399 studies from 10,467 records were included. The review reveals that there is an upward trend in mental illness detection NLP research. Deep learning methods receive more attention and perform better than traditional machine learning methods. We also provide some recommendations for future studies, including the development of novel detection methods, deep learning paradigms and interpretable models.

npj Digital Medicine (2022)5:46 ; https://doi.org/10.1038/s41746-022-00589-7

INTRODUCTION

Mental illnesses, also called mental health disorders, are highly prevalent worldwide, and have been one of the most serious public health concerns[1]. There are many different mental illnesses, including depression, suicidal ideation, bipolar disorder, autism spectrum disorder (ASD), anxiety disorder, schizophrenia, etc., any of which can have a negative influence on an individual's physical health and well-being with the problem exacerbated due to Covid-19[2]. According to the latest statistics, millions of people worldwide suffer from one or more mental disorders[1]. If mental illness is detected at an early stage, it can be beneficial to overall disease progression and treatment.

There are different text types, in which people express their mood, such as social media messages on social media platforms, transcripts of interviews and clinical notes including the description of patients' mental states. In recent years, natural language processing (NLP), a branch of artificial intelligence (AI) technologies, has played an essential role in supporting the analysis and management of large scale textual data and facilitating various tasks such as information extraction, sentiment analysis[3], emotion detection, and mental health surveillance[4–6]. Detecting mental illness from text can be cast as a text classification or sentiment analysis task, where we can leverage NLP techniques to automatically identify early indicators of mental illness to support early detection, prevention and treatment.

Existing reviews introduce mainly the computational methods for mental health illness detection, they mostly focus on specific mental illnesses (suicide[7–9], depression[10–12]), or specific data sources (social media[13–15], non-clinical texts[16]). To the best of our knowledge, there is no review of NLP techniques applied to mental illness detection from textual sources recently. We present a broader scope of mental illness detection using NLP that covers a decade of research, different types of mental illness and a variety of data sources. Our review aims to provide a comprehensive overview of the latest trends and recent NLP methodologies used for text-based mental illness detection, and also points at the future challenges and directions. Our review seeks to answer the following questions:

- What are the main NLP trends and approaches for mental illness detection?
- Which features have been used for mental health detection in traditional machine learning-based models?
- Which neural architectures have been commonly used to detect mental illness?
- What are the main challenges and future directions in NLP for mental illness?

SEARCH METHODOLOGY

Search strategy

A comprehensive search was conducted in multiple scientific databases for articles written in English and published between January 2012 and December 2021. The databases include PubMed, Scopus, Web of Science, DBLP computer science bibliography, IEEE Xplore, and ACM Digital Library.

The search query we used was based on four sets of keywords shown in Table 1. For mental illness, 15 terms were identified, related to general terms for mental health and disorders (e.g., mental disorder and mental health), and common specific mental illnesses (e.g., depression, suicide, anxiety). For data source, we searched for general terms about text types (e.g., social media, text, and notes) as well as for names of popular social media platforms, including Twitter and Reddit. The methods and detection sets refer to NLP methods used for mental illness identification.

The keywords of each sets were combined using Boolean operator "OR", and the four sets were combined using Boolean operator "AND". We conducted the searches in December 2021.

[1]Department of Computer Science, The University of Manchester, National Centre for Text Mining, Manchester, UK. [2]Department of Computer Science, Aalto University, Helsinki, Finland. [3]The Alan Turing Institute, London, UK. ✉email: sophia.ananiadou@manchester.ac.uk

Table 1. Keywords for literature search.

Category	Keywords
Mental illness (1)	Mental disorder, mental health, mental illness
	Depression, suicide, psychology, insomnia, stress, anxiety, schizophrenia, phobias, PTSD (post-traumatic stress disorder), ASD (autism spectrum disorder), anorexia, bipolar
Data sources (2)	Social media, text, language, posts, notes, interviews, records, survey
	Twitter, reddit, weibo, microblog, facebook, instagram
Methods (3)	Natural language processing, deep learning, machine learning, text mining, text analysis
	Neural network, CNN, LSTM, SVM, tree
Detection (4)	Detect, identify, recognize, predict, prevent, screen, assess, understand
Search query	(1) AND (2) AND (3) AND (4)

Filtering strategy

A total of 10,467 bibliographic records were retrieved from six databases, of which 7536 records were retained after removing duplication. Then, we used RobotAnalyst[17], a tool that minimizes the human workload involved in the screening phase of reviews, by prioritizing the most relevant articles for mental illness based on relevancy feedback and active learning[18,19].

Each of the 7536 records was screened based on title and abstract. Records were removed if the following exclusion criteria were met: (1) the full text was not available in English; (2) the abstract was not relevant to mental illness detection; (3) the method did not use textual experimental data, but speech or image data.

After the screening process, 611 records were retained for further review. An additional manual full-text review was conducted to retain only articles focusing on the description of NLP methods only. The final inclusion criteria were established as follow:

- Articles must study textual data such as contents from social media, electronic health records or transcription of interviews.
- They must focus on NLP methods for mental illness detection, including machine learning-based methods (in this paper, the machine learning methods refer to traditional feature engineering-based machine learning) and deep learning-based methods. We exclude review and data analysis papers.
- They must provide a methodology contribution by (1) proposing a new feature extraction method, a neural architecture, or a novel NLP pipeline; or (2) applying the learning methods to a specific mental health detection domain or task.

Following the full-text screening process, 399 articles were selected. The flow diagram of the article selection process is shown in Fig. 1.

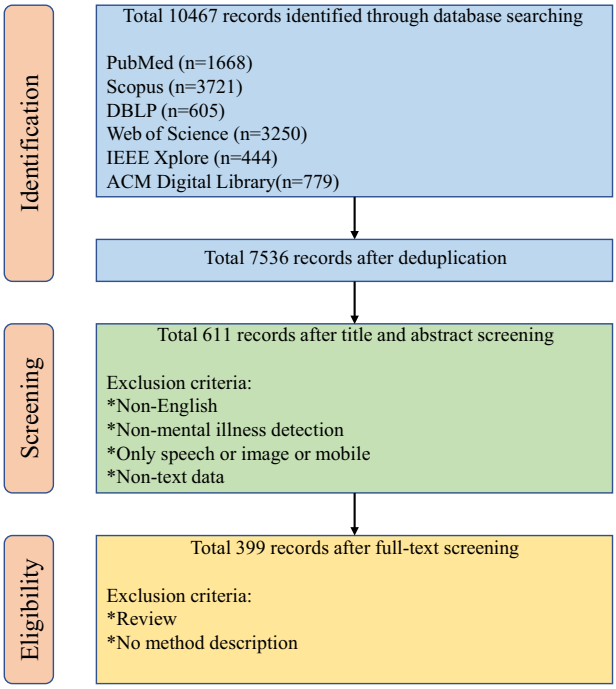

Fig. 1 Overview of article selection process. Six databases (PubMed, Scopus, Web of Science, DBLP computer science bibliography, IEEE Xplore, and ACM Digital Library) were searched. The flowchart lists reasons for excluding the study from the data extraction and quality assessment.

Data extraction

For each selected article, we extracted the following types of metadata and other information:

- Year of publication.
- The aim of research.
- The dataset used, including type of mental illness (e.g., depression, suicide, and eating disorder), language, and data sources (e.g., Twitter, electronic health records (EHRs) and interviews).
- The NLP method (e.g., machine learning and deep learning) and types of features used (e.g., semantic, syntactic, and topic).

FINDINGS

We show in Fig. 2 the number of publications retrieved and the methods used in our review, reflecting the trends of the past 10 years. We can observe that: (1) there is an upward trend in NLP-driven mental illness detection research, suggesting the great research value and prospects for automatic mental illness detection from text (2) deep learning-based methods have increased in popularity in the last couple of years.

In the following subsections, we provide an overview of the datasets and the methods used. In section Datesets, we introduce the different types of datasets, which include different mental illness applications, languages and sources. Section NLP methods used to extract data provides an overview of the approaches and summarizes the features for NLP development.

Datasets

In order to better train mental illness detection models, reliable and accurate datasets are necessary. There are several sources from which we can collect text data related to mental health, including social media posts, screening surveys, narrative writing, interviews and EHRs. At the same time, for different detection tasks, the datasets also differ in the types of illness

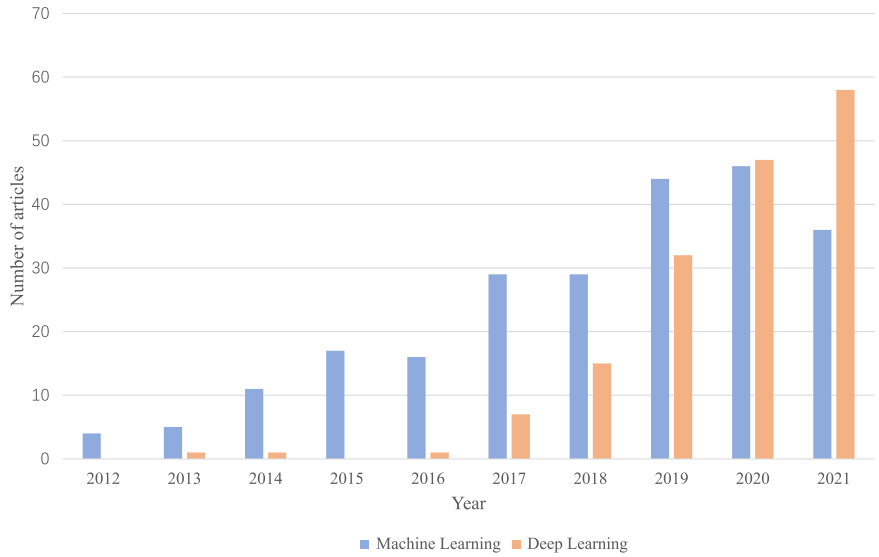

Fig. 2 NLP trends applied to mental illness detection research using machine learning and deep learning. The trend of the number of articles containing machine learning-based and deep learning-based methods for detecting mental illness from 2012 to 2021.

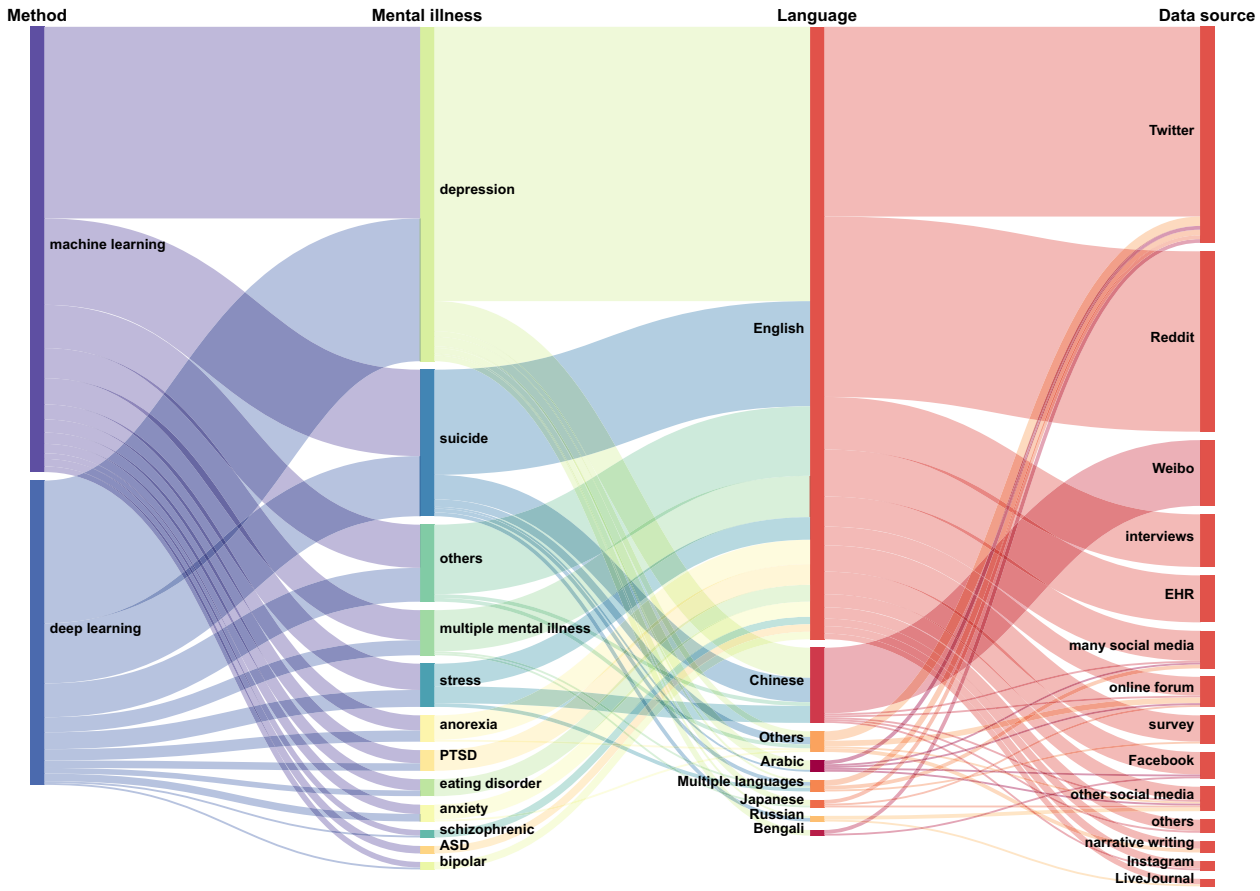

Fig. 3 Sankey diagram of NLP methods, illness, languages and applications. The different methods with their associated application are represented via flows. Nodes are represented as rectangles, and the height represents their value. The width of each curved line is proportional to their values.

they focus on and language. We show a comprehensive mapping of each method with its associated application using a Sankey diagram (Fig. 3).

Data sources. Figure 4 illustrates the distribution of the different data sources. It can be seen that, among the 399 reviewed papers, social media posts (81%) constitute the majority of sources,

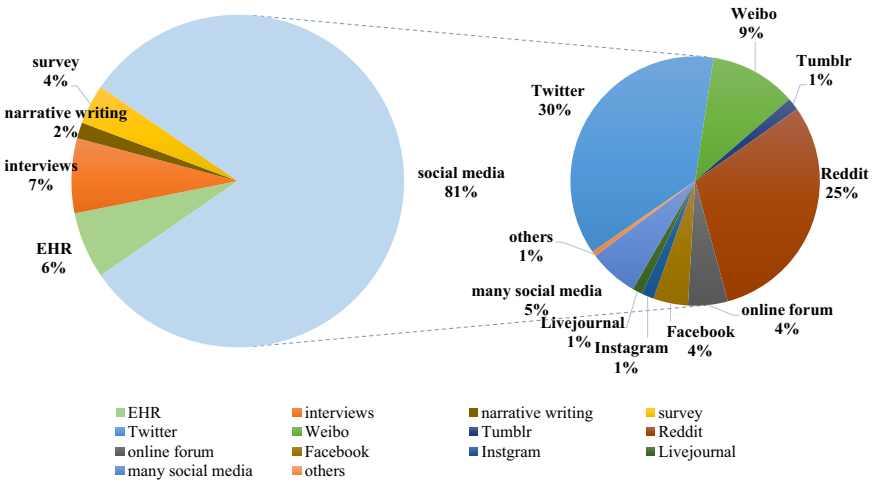

Fig. 4 Distribution of different data sources. The pie chart depicts the percentages of different textual data sources based on their numbers.

followed by interviews (7%), EHRs (6%), screening surveys (4%), and narrative writing (2%).

Social media posts

The use of social media has become increasingly popular for people to express their emotions and thoughts[20]. In addition, people with mental illness often share their mental states or discuss mental health issues with others through these platforms by posting text messages, photos, videos and other links. Prominent social media platforms are Twitter, Reddit, Tumblr, Chinese microblogs, and other online forums. We briefly introduce some popular social media platforms.

Twitter. Twitter is a popular social networking service with over 300 million active users monthly, in which users can post their tweets (the posts on Twitter) or retweet others' posts. Researchers can collect tweets using available Twitter application programming interfaces (API). For example, Sinha et al. created a manually annotated dataset to identify suicidal ideation in Twitter[21]. Hu et al. used a rule-based approach to label users' depression status from the Twitter[22]. However, normally Twitter does not allow the texts of downloaded tweets to be publicly shared, only the tweet identifiers—some/many of which may then disappear over time, so many datasets of actual tweets are not made publicly available[23].

Reddit. Reddit is also a popular social media platform for publishing posts and comments. The difference between Reddit and other data sources is that posts are grouped into different subreddits according to the topics (i.e., depression and suicide). Because of Reddit's open policy, their datasets are publicly available. Yates et al. established a depression dataset named "Reddit Self-reported Depression Diagnosis" (RSDD)[24], which contains about 9k depressed users and 100k control users. Similarly, CLEF risk 2019 shared task[25] also proposed an anorexia and self-harm detection task based on the Reddit platform.

Online forums. People can discuss their mental health conditions and seek mental help from online forums (also called online communities). There are various forms of online forums, such as chat rooms, discussion rooms (recoveryourlife, endthislife). For example, Saleem et al. designed a psychological distress detection model on 512 discussion threads downloaded from an online forum for veterans[26]. Franz et al. used the text data from TeenHelp.org, an Internet support forum, to train a self-harm detection system[27].

Electronic health records

EHRs, a rich source of secondary health care data, have been widely used to document patients' historical medical records[28]. EHRs often contain several different data types, including patients' profile information, medications, diagnosis history, images. In addition, most EHRs related to mental illness include clinical notes written in narrative form[29]. Therefore, it is appropriate to use NLP techniques to assist in disease diagnosis on EHRs datasets, such as suicide screening[30], depressive disorder identification[31], and mental condition prediction[32].

Interviews

Some work has been carried out to detect mental illness by interviewing users and then analyzing the linguistic information extracted from transcribed clinical interviews[33,34]. The main datasets include the DAIC-WoZ depression database[35] that involves transcriptions of 142 participants, the AViD-Corpus[36] with 48 participants, and the schizophrenic identification corpus[37] collected from 109 participants.

Screening surveys

In order to evaluate participants' mental health conditions, some researchers post questionnaires for clinician-patient diagnosis of patients or self-measurement. After participants are asked to fill in a survey from crowd-sourcing platforms (like Crowd Flower, Amazon's Mechanical Turk) or online platforms, the data is collected and labeled. There are different survey contents to measure different psychiatric symptoms. For depression, the PHQ-9 (Patient Health Questionnaire)[38] or Beck Depression Inventory (BDI) questionnaire[39] are widely used for assessing the severity of depressive symptoms. The Scale Center for Epidemiological Studies Depression Scale (CES-D) questionnaire[40] with 20 multiple-choice questions is also designed for testing depression. For suicide ideation, there are many questionnaires such as the Holmes-Rahe Social Readjustment Rating Scale (SRRS)[41] or the Depressive Symptom Inventory-Suicide Subscale (DSI-SS)[42].

Narrative writing

There are other types of texts written for specific experiments, as well as narrative texts that are not published on social media platforms, which we classify as narrative writing. For example, in one study, children were asked to write a story about a time that they had a problem or fought with other people, where researchers then analyzed their personal narrative to detect ASD[43]. In addition, a case study on Greek poetry of the 20th century was carried out for predicting suicidal tendencies[44].

Types of mental illness. There are many applications for the detection of different types of mental illness, where depression

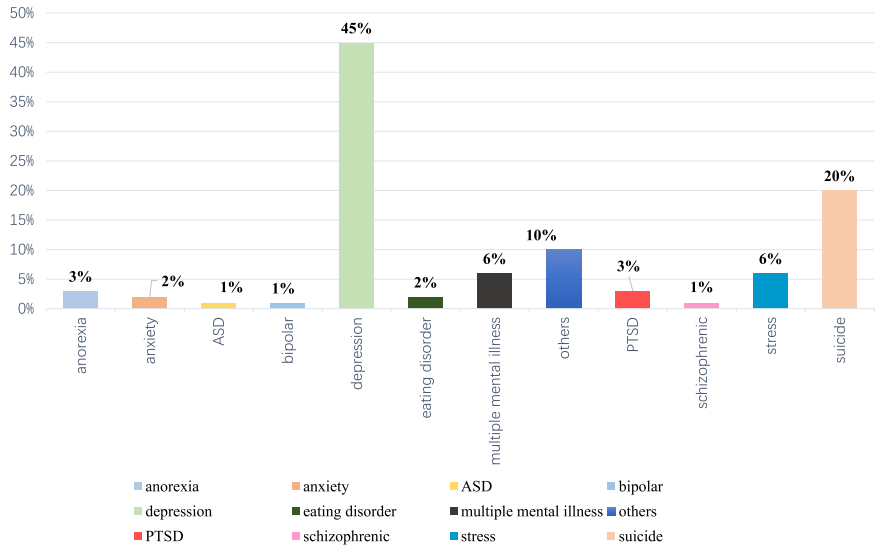

Fig. 5 Proportions of various types of mental illness. The chart depicts the percentages of different mental illness types based on their numbers.

(45%) and suicide (20%) account for the largest proportion; stress, anorexia, eating disorders, PTSD, bipolar disorder, anxiety, ASD and schizophrenia have corresponding datasets and have been analyzed using NLP (Fig. 5). This shows that there is a demand for NLP technology in different mental illness detection applications.

The amount of datasets in English dominates (81%), followed by datasets in Chinese (10%), Arabic (1.5%). When using non-English language datasets, the main difference lies in the pre-processing pipline, such as word segmentation, sentence splitting and other language-dependent text processing, while the methods and model architectures are language-agnostic.

NLP methods used to extract data

Machine learning methods. Traditional machine learning methods such as support vector machine (SVM), Adaptive Boosting (AdaBoost), Decision Trees, etc. have been used for NLP downstream tasks. Figure 3 shows that 59% of the methods used for mental illness detection are based on traditional machine learning, typically following a pipeline approach of data pre-processing, feature extraction, modeling, optimization, and evaluation.

In order to train a good ML model, it is important to select the main contributing features, which also help us to find the key predictors of illness. Table 2 shows an overview of commonly used features in machine learning. We further classify these features into linguistic features, statistical features, domain knowledge features, and other auxiliary features. The most frequently used features are mainly based on basic linguistic patterns (Part-of-Speech (POS)[45–47], Bag-of-words (BoW)[48–50], Linguistic Inquiry and Word Count (LIWC)[51–53]) and statistics (n-gram[54–56], term frequency-inverse document frequency (TF-IDF)[57–59], length of sentences or passages[60–62]) because these features can be easily obtained through text processing tools and are widely used in many NLP tasks. Furthermore, emotion and topic features have been shown empirically to be effective for mental illness detection[63–65]. Domain specific ontologies, dictionaries and social attributes in social networks also have the potential to improve accuracy[65–68]. Research conducted on social media data often leverages other auxiliary features to aid detection, such as social behavioral features[65,69], user's profile[70,71], or time features[72,73].

Machine learning models have been designed based on a combination of various extracted features. The majority of the papers based on machine learning methods used supervised learning, where they described one or more methods employed to detect mental illness: SVM[26,74–77], Adaptive Boosting (AdaBoost)[71,78–80], k-Nearest Neighbors (KNN)[38,81–83], Decision Tree[84–87], Random Forest[75,88–90], Logistic Model Tree (LMT)[47,47,91,92], Naive Bayes (NB)[64,86,93,94], Logistic Regression[37,95–97], XGBoost[38,55,98,99], and some ensemble models combining several methods[75,100–102]. The advantage of such supervised learning lies in the model's ability to learn patterns from labeled data, thus ensuring better performance. However, labeling the large amount of data at a high quality level is time-consuming and challenging, although there are methods that help reduce the human annotation burden[103]. Thus, we need to use other methods which do not rely on labeled data or need only a small amount of data to train a classifier.

Unsupervised learning methods to discover patterns from unlabeled data, such as clustering data[55,104,105], or by using LDA topic model[27]. However, in most cases, we can apply these unsupervised models to extract additional features for developing supervised learning classifiers[56,85,106,107]. Across all papers, few papers[108,109] used semi-supervised learning (models trained from large unlabeled data as additional information), including statistical model ssToT (semi-supervised topic modeling over time)[108] and classic semi-supervised algorithms (YATSI[110] and LLGC[111]).

Deep learning methods. As mentioned above, machine learning-based models rely heavily on feature engineering and feature extraction. Using deep learning frameworks allows models to capture valuable features automatically without feature engineering, which helps achieve notable improvements[112]. Advances in deep learning methods have brought breakthroughs in many fields including computer vision[113], NLP[114], and signal processing[115]. For the task of mental illness detection from text, deep learning techniques have recently attracted more attention and shown better performance compared to machine learning ones[116].

Deep learning-based frameworks mainly contain two layers: an embedding layer and a classification layer. By using an embedding layer, the inputs are embedded from sparse one-hot encoded vectors (where only one member of a vector is '1', all others are '0', leading to the sparsity) into dense vectors which can preserve semantic and syntactic information such that deep learning models can be better trained[117]. There are many different embedding techniques, such as ELMo, GloVe word embedding[118], word2vec[119] and contextual language encoder representations (e.g., bidirectional encoder representations from transformers (BERT)[120] and ALBERT[121]).

Table 2. An overview of features used in machine learning-based models.

Feature categories	Feature types	Features	Description	Typical examples
Linguistic features	Sysntactic features	Part-of-Speech (POS)	Based on the grammatical use and functions, words are categorized into different types of POS (like None, Verb, Adverb).	45–47
		Dependency parsing	The grammatical structure of a sentence.	205,206
	Lexicon-based features	Bag-of-words (BoW)	The simplest form of text representation using numbers of vocabularies.	48–50
		Lexical diversity, lexical density	The unique vocabulary usage and proportion of content words.	37
	Emotion features	Sentiment scores	Sentiment scores are used to quantify the feeling of texts and determine the sentiment polarity (positive, negtive,or netural). The way of calculation includes using VADER sentiment analysis (Valence Aware Dictionary and sEntiment Reasoner)[207], SenticNet 5 lexicon[208], AFINN lexicon[209].	63,210–212
		Emotion scores	The emotion scores indicates the user's emotions and opinions of texts to an extent, which is beneficial for mental issues detection. NRC Affect Intensity Lexicon[213] are always used.	56, 63, 109,214
	Semantic features	Semantic similarity	Using semantic similarity predict whether the sentence or word is sematically related to the target sentence or word.	60,215
	Topic features	Topic features	The topics extracted from texts using some topic-modeling algorithms, like Latent Dirichlet Allocation (LDA)[216], Latent Semantic Analysis (LSA)[217], Non-negative matrix factorization (NMF)[218].	55, 65, 87,219
	Linguistic features	LIWC	Linguistic Inquiry and Word Count (LIWC)[220] is always used to automatically extract linguistic styles from texts by calculating the percentages of words in different categories, like linguistic, social affective, etc..	51–53,82
	Others	Hashtag, emoji	Hashtag is metadata tag from social media platform,which present a theme or topic. emoticons or emojis are often used to show various types of emotions.	78, 79,221
Statistical features	Statistical corpus features	n-gram	N-gram is a contiguous sequence of n words.	54–56
		TF-IDF	Term frequency-inverse document frequency (TF-IDF) reflect the importance of the word in document.	57–59,222
		Length statistics	The length of posts, documents or average sentence.	60–62,223
	Vector-based features	Word embedding	The vector-based representation of words. Examples: word2vec[224], GloVe[118].	49, 56, 106,225
		Document embedding	The vector-based representation of document.	226
Domain knowledge features	Conceptual features	UMLS	Unified Medical Language System (UMLS) is a set of key terminology, coding standards, and associated resources related to biomedical information.	67,227
		Linguistic dictionary	The dictionary contains mental health illness related words	66, 228,229
Other auxiliary features	Social behavioral features	Social connectivity	The degree of social interaction on social media, like number of followers, friends, and communities joined[230].	68
		User behaviors	The user's behavioral signals on social media, such as the frequency of comments and forwardings.	65, 69,231
	Time features	Time features	Focusing on the time-related features, like sending time, time interval.	72,73
	User's profile features	User's profile features	The user's profile features contain their individual information on social networks.	70, 71,231

Table 3. The deep learning methods for mental illness detection.

Type	Method	Description
CNN-based methods	Standard CNN[122–127]	Standard CNN structure: convolutional layer, pooling layer and fully connected layer. Some studies also incorporate other textual features (like POS, LIWC, BoW, etc.).
	Multi-Gated LeakyReLU CNN (MGL-CNN)[128]	Two hierarchical (post-level and user-level)neural network models with gated units and convolutional networks.
	Graph model combined with Convolutional Neural Network[129]	A unified hybrid model combining CNN with factor graph model which leverages social interactions and content.
RNN-based methods	LSTM or GRU (some with attention mechanism)[32,133,136,232–234]	Standard RNN structure: Long Short-Term Memory networks (LSTM) or Gate Recurrent Unit(GRU), and some studies add attention mechanism.
	Hierarchical Attention Network (HAN) with GRU[138]	The GRU with a word-level attention layer and a sentence-level attention layer.
	LSTM with transfer learning[140,141]	Using transfer learning on open dataset for model pre-training.
	LSTM or GRU with multi-task learning[142,235–237]	Using multi-task learning to help illness detection get better result. The tasks include multi-risky behaviors classification, severity score prediction,word vector classification,and sentiment classification.
	LSTM or GRU with reinforcement learning[143,144]	Using reinforcement learning to automatically select the important posts.
	LSTM or GRU with multiple instance learning[145,146]	Using multiple instance learning to get the possibility of post-level labels and improve the prediction of user-level labels.
	SISMO[139]	An ordinal hierarchical LSTM attention model
Transformer-based methods	Self-attention models[148,149]	Using the encoder structure of transformer which has self-attention module.
	BERT-based models (BERT[150,151], DistilBERT[152], RoBERTa[153], ALBERT[150], BioClinical BERT[31], XLNET[154], GPT-1[155])	Different BERT-based pre-trained models.
Hybrid-based methods	LSTM+CNN[156–160]	Combining LSTM with CNN to extract local features and sequence features.
	STATENet (using transformer and LSTM)[161]	A time-aware transformer combining emotional and historical information.
	Sub-emotion network[164,165,238]	Integrating Bag-of-Sub-Emotion embeddings into LSTM to get emotional information.
	Events and Personality traits for Stress Prediction (EPSP) model[239]	A joint memory network for learning the dynamics of user's emotions and personality.
	PHASE[166]	A time and phase-aware model that learns historical emotional features from users.
	Hyperbolic graph convolution networks[167]	Hyperbolic Graph Convolutions with the Hawkes process to learn the historical emotional spectrum of a user.

According to the structures of different classification layer's structures, we have divided the deep learning-based methods into the following categories for this review: convolutional neural networks (CNN)-based methods (17%), recurrent neural networks (RNN)-based methods (36%), transformer-based methods (17%) and hybrid-based methods (30%) that combine multiple neural networks with different structures, as shown in Table 3.

CNN-based methods. The standard CNN structure is composed of a convolutional layer and a pooling layer, followed by a fully-connected layer. Some studies[122–127] utilized standard CNN to construct classification models, and combined other features such as LIWC, TF-IDF, BOW, and POS. In order to capture sentiment information, Rao et al. proposed a hierarchical MGL-CNN model based on CNN[128]. Lin et al. designed a CNN framework combined with a graph model to leverage tweet content and social interaction information[129].

RNN-based methods. The architecture of RNNs allows previous outputs to be used as inputs, which is beneficial when using sequential data such as text. Generally, long short-term memory (LSTM)[130] and gated recurrent (GRU)[131] networks models that can deal with the vanishing gradient problem[132] of the traditional RNN are effectively used in NLP field. There are many studies (e.g.,[133,134]) based on LSTM or GRU, and some of them[135,136] exploited an attention mechanism[137] to find significant word information from text. Some also used a hierarchical attention network based on LSTM or GRU structure to better exploit the different-level semantic information[138,139].

Moreover, many other deep learning strategies are introduced, including transfer learning, multi-task learning, reinforcement learning and multiple instance learning (MIL). Rutowski et al. made use of transfer learning to pre-train a model on an open dataset, and the results illustrated the effectiveness of pre-training[140,141]. Ghosh et al. developed a deep multi-task method[142] that modeled emotion recognition as a primary task and depression detection as a secondary task. The experimental results showed that multi-task frameworks can improve the performance of all tasks when jointly learning. Reinforcement learning was also used in depression detection[143,144] to enable the model to pay more attention to useful information rather than noisy data by selecting indicator posts. MIL is a machine learning paradigm, which aims to learn features from bags' labels of the training set instead of individual labels. Wongkoblap et al. used MIL to predict users with depression task[145,146].

Transformer-based methods. Recently, transformer architectures[147] were able to solve long-range dependencies using attention and recurrence. Wang et al. proposed the C-Attention

network[148] by using a transformer encoder block with multi-head self-attention and convolution processing. Zhang et al. also presented their TransformerRNN with multi-head self-attention[149]. Additionally, many researchers leveraged transformer-based pre-trained language representation models, including BERT[150,151], DistilBERT[152], Roberta[153], ALBERT[150], BioClinical BERT for clinical notes[31], XLNET[154], and GPT model[155]. The usage and development of these BERT-based models prove the potential value of large-scale pre-training models in the application of mental illness detection.

Hybrid-based methods. Some methods combining several neural networks for mental illness detection have been used. For examples, the hybrid frameworks of CNN and LSTM models [156–160] are able to obtain both local features and long-dependency features, which outperform the individual CNN or LSTM classifiers used individually. Sawhney et al. proposed STATENet[161], a time-aware model, which contains an individual tweet transformer and a Plutchik-based emotion[162] transformer to jointly learn the linguistic and emotional patterns. Inspired by the improved performance of using sub-emotions representations[163], Aragon et al. presented a deep emotion attention model[164] which consists of sub-emotion embedding, CNN, GRU as well as an attention mechanism, and Lara et al. also proposed Deep Bag of Sub-Emotions (DeepBose) model[165]. Furthermore, Sawhney et al. introduced the PHASE model[166], which learns the chronological emotional progression of a user by a new time-sensitive emotion LSTM and also Hyperbolic Graph Convolution Networks[167]. It also learns the chronological emotional spectrum of a user by using BERT fine-tuned for emotions as well as a heterogeneous social network graph.

Evaluation metrics. Evaluation metrics are used to compare the performance of different models for mental illness detection tasks. Some tasks can be regarded as a classification problem, thus the most widely used standard evaluation metrics are Accuracy (AC), Precision (P), Recall (R), and F1-score (F1)[149,168–170]. Similarly, the area under the ROC curve (AUC-ROC)[60,171,172] is also used as a classification metric which can measure the true positive rate and false positive rate. In some studies, they can not only detect mental illness, but also score its severity[122,139,155,173]. Therefore, metrics of mean error (e.g., mean absolute error, mean square error, root mean squared error)[173] and other new metrics (e.g., graded precision, graded recall, average hit rate, average closeness rate, average difference between overall depression levels)[139,174] are sometimes needed to indicate the difference between the predicted severity and the actual severity in a dataset. Meanwhile, taking into account the timeliness of mental illness detection, where early detection is significant for early prevention, an error metric called early risk detection error was proposed[175] to measure the delay in decision.

DISCUSSION

Although promising results have been obtained using both machine and deep learning methods, several challenges remain for the mental illness detection task that require further research. Herein, we introduce some key challenges and future research directions:

- Data volume and quality: Most of the methods covered in this review used supervised learning models. The success of these methods is attributed to the number of training datasets available. These training datasets often require human annotation, which is usually a time-consuming and expensive process. However, in the mental illness detection task, there are not enough annotated public datasets. For training reliable models, the quality of datasets is concerning. Some datasets have annotation bias because the annotators can not confirm a definitive action has taken place in relation to a disorder (e.g., if actual suicide has occurred) and can only label them within the constraints of their predefined annotation rules[9]. In addition, some imbalanced datasets have many negative instances (individuals without mental disorders), which is not conducive to training comprehensive and robust models. Therefore, it is important to explore how to train a detection model by using a small quantity of labeled training data or not using training data. Semi-supervised learning[176] incorporates few labeled data and large amounts of unlabeled data into the training process, which can be used to facilitate annotation[177] or improve classification performance when labeled data is scarce. Additionally, unsupervised methods can also be applied in mental disorders detection. For instance, unsupervised topic modeling[178] increases the explainability of results and aids the extraction of latent features for developing further supervised models.[179,180]
- Performance and instability: There are some causes of model instability, including class imbalance, noisy labels, and extremely long or extremely short text samples text. Performance is not robust when training on the datasets from different data sources due to diverse writing styles and semantic heterogeneity. Thus, the performance of some detection models is not good. With the advances of deep learning techniques, various learning techniques have emerged and accelerated NLP research, such as adversarial training[181], contrastive learning[182], joint learning[183], reinforcement learning[184] and transfer learning[185], which can also be utilized for the mental illness detection task. For example, pre-trained Transformer-based models can be transferred to anorexia detection in Spanish[186], and reinforcement networks can be used to find the sentence that best reflects the mental state. Other emerging techniques like attention mechanism[187], knowledge graph[188], and commonsense reasoning[189], can also be introduced for textual feature extraction. In addition, feature enrichment and data augmentation[190] are useful to achieve comparable results. For example, many studies use multi-modal data resources, such as image[191–193], and audio[194–196], which perform better than the single-modal text-based model.
- Interpretability: The goal of representation learning for mental health is to understand the causes or explanatory factors of mental illness in order to boost detection performance and empower decision-making. The evaluation of a successful model does not only rely on performance, but also on its interpretability[197], which is significant for guiding clinicians to understand not only what has been extracted from text but the reasoning underlying some prediction[198–200]. Deep learning-based methods achieve good performance by utilizing feature extraction and complex neural network structures for illness detection. Nevertheless, they are still treated as black boxes[201] and fail to explain the predictions. Therefore, in future work, the explainability of the deep learning models will become an important research direction.
- Ethical considerations: It is of greater importance to discuss ethical concerns when using mental health-related textual data, since the privacy and security of personal data is significant and health data is particularly sensitive. During the research, the researchers should follow strict protocols similar to the guidelines[202] introduced by Bentan et al., to ensure the data is properly applied in healthcare research while protecting privacy to avoid further psychological distress. Furthermore, when using some publicly available data, researchers need to acquire ethical approvals from institutional review boards and human research ethics committees[203,204].

There has been growing research interest in the detection of mental illness from text. Early detection of mental disorders is an

important and effective way to improve mental health diagnosis. In our review, we report the latest research trends, cover different data sources and illness types, and summarize existing machine learning methods and deep learning methods used on this task.

We find that there are many applications for different data sources, mental illnesses, even languages, which shows the importance and value of the task. Our findings also indicate that deep learning methods now receive more attention and perform better than traditional machine learning methods.

We discuss some challenges and propose some future directions. In the future, the development of new methods including different learning strategies, novel deep learning paradigms, interpretable models and multi-modal methods will support mental illness detection, with an emphasis on interpretability being crucial for uptake of detection applications by clinicians.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: 26 October 2021; Accepted: 23 February 2022;
Published online: 08 April 2022

REFERENCES

1. Rehm, J. & Shield, K. D. Global burden of disease and the impact of mental and addictive disorders. *Curr. Psychiatry Rep.* **21**, 1–7 (2019).
2. Santomauro, D. F. et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet* **398**, 1700–1712 (2021).
3. Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. Natural language processing: an introduction. *J. Am. Med. Inform. Assoc.* **18**, 544–551 (2011).
4. Ive, J. Generation and evaluation of artificial mental health records for natural language processing. *NPJ Digital Med.* **3**, 1–9 (2020).
5. Mukherjee, S. S. et al. Natural language processing-based quantification of the mental state of psychiatric patients. *Comput. Psychiatry* **4**, 76–106 (2020).
6. Jackson, R. G. Natural language processing to extract symptoms of severe mental illness from clinical text: the clinical record interactive search comprehensive data extraction (cris-code) project. *BMJ Open* **7**, 012012 (2017).
7. Castillo-Sánchez, G. Suicide risk assessment using machine learning and social networks: a scoping review. *J. Med. Syst.* **44**, 1–15 (2020).
8. Franco-Martín, M. A. A systematic literature review of technologies for suicidal behavior prevention. *J. Med. Syst.* **42**, 1–7 (2018).
9. Ji, S. Suicidal ideation detection: a review of machine learning methods and applications. *IEEE Trans. Comput. Soc. Syst.* **8**, 214–226 (2021).
10. Giuntini, F. T. A review on recognizing depression in social networks: challenges and opportunities. *J. Ambient Intell. Human. Comput.* **11**, 4713–4729 (2020).
11. Mahdy, N., Magdi, D. A., Dahroug, A. & Rizka, M. A. Comparative study: different techniques to detect depression using social media. in Internet of Things-Applications and Future, pp. 441–452 (2020).
12. Khan, A., Husain, M. S. & Khan, A. Analysis of mental state of users using social media to predict depression! a survey. *Int. J. Adv. Res. Comput. Sci.* **9**, 100–106 (2018).
13. Skaik, R. & Inkpen, D. Using social media for mental health surveillance: a review. *ACM Comput. Surv.* **53**, 1–31 (2020).
14. Chancellor, S. & De Choudhury, M. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Med.* **3**, 1–11 (2020).
15. Ríssola, E. A., Losada, D. E. & Crestani, F. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthc.* **2**, 1–31 (2021).
16. Calvo, R. A., Milne, D. N., Hussain, M. S. & Christensen, H. Natural language processing in mental health applications using non-clinical texts. *Nat. Lang. Eng.* **23**, 649–685 (2017).
17. Przybyła, P. Prioritising references for systematic reviews with robotanalyst: a user study. *Res. Synth. Methods* **9**, 470–488 (2018).
18. O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. & Ananiadou, S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst. Rev.* **4**, 1–22 (2015).
19. Miwa, M., Thomas, J., O'Mara-Eves, A. & Ananiadou, S. Reducing systematic review workload through certainty-based screening. *J. Biomed. Inform.* **51**, 242–253 (2014).
20. Kemp, S. Digital 2020: 3.8 billion people use social media. *We Are Social* **30**, (2020). https://wearesocial.com/uk/blog/2020/01/digital-2020-3-8-billion-people-use-social-media/.
21. Sinha, P. P. et al. Suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In *Proc. 28th ACM International Conference on Information and Knowledge Management*, pp. 941–950 (2019).
22. Hu, P. et al. Bluememo: depression analysis through twitter posts. In *IJCAI*, pp. 5252–5254 (2020).
23. Golder, S., Ahmed, S., Norman, G. & Booth, A. Attitudes toward the ethics of research using social media: a systematic review. *J. Med. Internet Res.* **19**, 7082 (2017).
24. Yates, A., Cohan, A. & Goharian, N. Depression and self-harm risk assessment in online forums. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing* (2017).
25. Naderi, N., Gobeill, J., Teodoro, D., Pasche, E. & Ruch, P. A baseline approach for early detection of signs of anorexia and self-harm in reddit posts. In *CLEF (Working Notes)* (2019).
26. Saleem, S. et al. Automatic detection of psychological distress indicators in online forum posts. In *Proc. 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–4 (2012).
27. Franz, P. J., Nook, E. C., Mair, P. & Nock, M. K. Using topic modeling to detect and describe self-injurious and related content on a large-scale digital platform. *Suicide Life Threat. Behav.* **50**, 5–18 (2020).
28. Menachemi, N. & Collum, T. H. Benefits and drawbacks of electronic health record systems. *Risk Manag. Healthc. Policy* **4**, 47 (2011).
29. Kho, A. N. Practical challenges in integrating genomic data into the electronic health record. *Genet. Med.* **15**, 772–778 (2013).
30. Downs, J. et al. Detection of suicidality in adolescents with autism spectrum disorders: developing a natural language processing approach for use in electronic health records. In *AMIA Annual Symposium Proceedings*, vol. 2017, p. 641 (2017).
31. Kshatriya, B. S. A. et al. Neural language models with distant supervision to identify major depressive disorder from clinical notes. Preprint at *arXiv* https://arxiv.org/abs/2104.09644 (2021).
32. Tran, T. & Kavuluru, R. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. *J. Biomed. Inform.* **75**, 138–148 (2017).
33. Morales, M. R. & Levitan, R. Speech vs. text: a comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 136–143 (2016).
34. Arseniev-Koehler, A., Mozgai, S. & Scherer, S. What type of happiness are you looking for?-a closer look at detecting mental health from language. In *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 1–12 (2018).
35. Ringeval, F. et al. Avec 2017: real-life depression, and affect recognition workshop and challenge. In *Proc. 7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 3–9 (2017).
36. Valstar, M. et al. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proc. 4th International Workshop on Audio/visual Emotion Challenge*, pp. 3–10 (2014).
37. Voleti, R. et al. Objective assessment of social skills using automated language analysis for identification of schizophrenia and bipolar disorder. In *Proc. Interspeech*, pp. 1433–1437 (2019).
38. Tlachac, M., Toto, E. & Rundensteiner, E. You're making me depressed: Leveraging texts from contact subsets to predict depression. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 1–4 (2019).
39. Stankevich, M., Smirnov, I., Kiselnikova, N. & Ushakova, A. Depression detection from social media profiles. In *International Conference on Data Analytics and Management in Data Intensive Domains*, pp. 181–194 (2019).
40. Wongkoblap, A., Vadillo, M. A. & Curcin, V. A multilevel predictive model for detecting social network users with depression. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 130–135 (2018).
41. Delgado-Gomez, D., Blasco-Fontecilla, H., Sukno, F., Ramos-Plasencia, M. S. & Baca-Garcia, E. Suicide attempters classification: toward predictive models of suicidal behavior. *Neurocomputing* **92**, 3–8 (2012).
42. von Glischinski, M., Teismann, T., Prinz, S., Gebauer, J. E. & Hirschfeld, G. Depressive symptom inventory suicidality subscale: optimal cut points for clinical and non-clinical samples. *Clin. Psychol. Psychother.* **23**, 543–549 (2016).

43. Hilvert, E., Davidson, D. & Gámez, P. B. Assessment of personal narrative writing in children with and without autism spectrum disorder. *Res. Autism Spectr. Disord.* **69**, 101453 (2020).
44. Zervopoulos, A. D. et al. Language processing for predicting suicidal tendencies: a case study in greek poetry. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 173–183 (2019).
45. Birjali, M., Beni-Hssane, A. & Erritali, M. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Proc. Computer. Sci.* **113**, 65–72 (2017).
46. Trifan, A., Antunes, R., Matos, S. & Oliveira, J. L. Understanding depression from psycholinguistic patterns in social media texts. *Adv. Inf. Retr.* **12036**, 402 (2020).
47. Briand, A., Almeida, H. & Meurs, M. -J. Analysis of social media posts for early detection of mental health conditions. In *Canadian Conference on Artificial Intelligence*, pp. 133–143 (2018).
48. Trifan, A. & Oliveira, J. L. Bioinfo@ uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders. In *CLEF (Working Notes)* (2019).
49. Lin, W., Ji, D. & Lu, Y. Disorder recognition in clinical texts using multi-label structured svm. *BMC Bioinform.* **18**, 1–11 (2017).
50. Chomutare, T. Text classification to automatically identify online patients vulnerable to depression. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pp. 125–130 (2014).
51. Islam, M. R. Depression detection from social network data using machine learning techniques. *Health Inf. Sci. Syst.* **6**, 1–12 (2018).
52. Su, Y., Zheng, H., Liu, X. & Zhu, T. Depressive emotion recognition based on behavioral data. In *International Conference on Human Centered Computing*, pp. 257–268 (2018).
53. Simms, T. et al. Detecting cognitive distortions through machine learning text analytics. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 508–512 (2017).
54. He, Q., Veldkamp, B. P., Glas, C. A. & de Vries, T. Automated assessment of patients' self-narratives for posttraumatic stress disorder screening using natural language processing and text mining. *Assessment* **24**, 157–172 (2017).
55. Shickel, B., Siegel, S., Heesacker, M., Benton, S. & Rashidi, P. Automatic detection and classification of cognitive distortions in mental health text. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 275–280 (2020).
56. Guntuku, S. C., Giorgi, S. & Ungar, L. Current and future psychological health prediction using language and socio-demographics of children for the clpysch 2018 shared task. In *Proc. Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 98–106 (2018).
57. Stankevich, M., Isakov, V., Devyatkin, D. & Smirnov, I. V. Feature engineering for depression detection in social media. In *ICPRAM*, pp. 426–431 (2018).
58. Boag, W. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Transl. Psychiatry* **11**, 1–6 (2021).
59. Adamou, M. et al. Mining free-text medical notes for suicide risk assessment. In *Proc. 10th Hellenic Conference on Artificial Intelligence*, pp. 1–8 (2018).
60. Saleem, S. et al. Automatic detection of psychological distress indicators and severity assessment from online forum posts. In *Proc. COLING 2012*, pp. 2375–2388 (2012).
61. Trifan, A. & Oliveira, J. L. Cross-evaluation of social mining for classification of depressed online personas. *J. Integr. Bioinform.* (2021)
62. Balani, S. & De Choudhury, M. Detecting and characterizing mental health related self-disclosure in social media. In *Proc. 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 1373–1378 (2015).
63. Delahunty, F., Wood, I. D. & Arcan, M. First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot. In *Irish Conference on Artificial Intelligence and Cognitive Science* (2018).
64. Deshpande, M. & Rao, V. Depression detection using emotion artificial intelligence. In *2017 International Conference on Intelligent Sustainable Systems (iciss)*, pp. 858–862 (2017).
65. Hwang, Y., Kim, H. J., Choi, H. J. & Lee, J. Exploring abnormal behavior patterns of online users with emotional eating behavior: topic modeling study. *J. Med. Internet Res.* **22**, 15700 (2020).
66. Alam, M. A. U. & Kapadia, D. Laxary: a trustworthy explainable twitter analysis model for post-traumatic stress disorder assessment. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 308–313 (2020).
67. Plaza-del Arco, F. M., López-Úbeda, P., Diaz-Galiano, M. C., Urena-López, L. A. & Martin-Valdivia, M.-T. *Integrating Umls for Early Detection of Sings of Anorexia*. (Universidad de Jaen, Campus Las Lagunillas: Jaen, Spain, 2019).
68. Dao, B., Nguyen, T., Phung, D. & Venkatesh, S. Effect of mood, social connectivity and age in online depression community via topic and linguistic analysis. In *International Conference on Web Information Systems Engineering*, pp. 398–407 (2014).
69. Katchapakirin, K., Wongpatikaseree, K., Yomaboot, P. & Kaewpitakkun, Y. Facebook social media for depression detection in the thai community. In *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 1–6 (2018).
70. Chang, M. -Y. & Tseng, C. -Y. Detecting social anxiety with online social network data. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pp. 333–336 (2020).
71. Tong. L. et al. Cost-sensitive boosting pruning trees for depression detection on Twitter. In *IEEE Transactions on Affective Computing*, https://doi.org/10.1109/TAFFC.2022.3145634 (2019).
72. Guntuku, S. C., Buffone, A., Jaidka, K., Eichstaedt, J. C. & Ungar, L. H. Understanding and measuring psychological stress using social media. In *Proc. International AAAI Conference on Web and Social Media*, vol. 13, pp. 214–225 (2019).
73. Zhao, L., Jia, J. & Feng, L. Teenagers' stress detection based on time-sensitive micro-blog comment/response actions. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pp. 26–36 (2015).
74. Ziwei, B. Y. & Chua, H. N. An application for classifying depression in tweets. In *Proc. 2nd International Conference on Computing and Big Data*, pp. 37–41 (2019).
75. Prakash, A., Agarwal, K., Shekhar, S., Mutreja, T. & Chakraborty, P. S. An ensemble learning approach for the detection of depression and mental illness over twitter data. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 565–570 (2021).
76. Coello-Guilarte, L., Ortega-Mendoza, R. M., Villaseñor-Pineda, L. & Montes-y-Gómez, M. Crosslingual depression detection in twitter using bilingual word alignments. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 49–61 (2019).
77. Qiu, J. & Gao, J. Depression tendency recognition model based on college student's microblog text. In *International Conference on Intelligence Science*, pp. 351–359 (2017).
78. Almouzini, S. et al. Detecting arabic depressed users from twitter data. *Proc. Comput. Sci.* **163**, 257–265 (2019).
79. Mbarek, A., Jamoussi, S., Charfi, A. & Hamadou, A. B. Suicidal profiles detection in twitter. In *WEBIST*, pp. 289–296 (2019).
80. Xu, S. et al. Automatic verbal analysis of interviews with schizophrenic patients. In *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pp. 1–5 (2018).
81. Verma, P., Sharma, K. & Walia, G. S. Depression detection among social media users using machine learning. In *International Conference on Innovative Computing and Communications*, pp. 865–874 (2021).
82. Shrestha, A. & Spezzano, F. Detecting depressed users in online forums. In *Proc. 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 945–951 (2019).
83. Desmet, B. & Hoste, V. Recognising suicidal messages in dutch social media. In *9th International Conference on Language Resources and Evaluation (LREC)*, pp. 830–835 (2014).
84. He, L. & Luo, J. "What makes a pro eating disorder hashtag": using hashtags to identify pro eating disorder tumblr posts and twitter users. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3977–3979 (2016).
85. Marerngsit, S. & Thammaboosadee, S. A two-stage text-to-emotion depressive disorder screening assistance based on contents from online community. In *2020 8th International Electrical Engineering Congress (iEECON)*, pp. 1–4 (2020).
86. Nadeem, M. Identifying depression on twitter. Preprint at *arXiv* https://arxiv.org/abs/1607.07384 (2016).
87. Fodeh, S. et al. Using machine learning algorithms to detect suicide risk factors on twitter. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pp. 941–948 (2019).
88. Tariq, S. A novel co-training-based approach for the classification of mental illnesses using social media posts. *IEEE Access* **7**, 166165–166172 (2019).
89. Mittal, A., Goyal, A. & Mittal, M. Data preprocessing based connecting suicidal and help-seeking behaviours. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1824–1830 (2021).
90. Kamite, S. R. & Kamble, V. Detection of depression in social media via twitter using machine learning approach. In *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)*, pp. 122–125 (2020).
91. Schoene, A. M. & Dethlefs, N. Automatic identification of suicide notes from linguistic and sentiment features. In *Proc. 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 128–133 (2016).
92. Almeida, H., Briand, A. & Meurs, M.- J. Detecting early risk of depression from social media user-generated content. In *CLEF* (Working Notes) (2017).
93. Govindasamy, K. A. & Palanichamy, N. Depression detection using machine learning techniques on twitter data. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 960–966 (2021).
94. Baheti, R. & Kinariwala, S. Detection and analysis of stress using machine learning techniques. *Int. J. Eng. Adv. Technol.* **9**, 335–342 (2019).

95. Németh, R., Sik, D. & Máté, F. Machine learning of concepts hard even for humans: the case of online depression forums. *Int. J. Qualitative Methods* **19**, 1609406920949338 (2020).
96. Benton, A., Mitchell, M. & Hovy, D. Multitask learning for mental health conditions with limited social media data. In *Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 1*, pp. 152–162 (2017).
97. Hiraga, M. Predicting depression for japanese blog text. In *Proc. ACL 2017, Student Research Workshop*, pp. 107–113 (2017).
98. Nasir, A., A slam, K., Tariq, S. & Ullah, M. F. Predicting mental illness using social media posts and comments. *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. **11** (2020).
99. Skaik, R. & Inkpen, D. Using twitter social media for depression detection in the canadian population. In *2020 3rd Artificial Intelligence and Cloud Computing Conference*, pp. 109–114 (2020).
100. Chadha, A. & Kaushik, B. Machine learning based dataset for finding suicidal ideation on twitter. In *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 823–828 (2021).
101. Sekulić, I., Gjurković, M. & Šnajder, J. Not Just Depressed: Bipolar Disorder Prediction on Reddit. In *Proc, the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 72–78 (2018).
102. Kumar, A., Sharma, A. & Arora, A. Anxious depression prediction in real-time social data. In *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019*. (Uttaranchal University, Dehradun, India, 2019).
103. Nghiem, M. -Q., Baylis, P. & Ananiadou, S. Paladin: an annotation tool based on active and proactive learning. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 238–243 (2021).
104. Park, A., Conway, M. & Chen, A. T. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. *Comput. Hum. Behav.* **78**, 98–112 (2018).
105. Shrestha, A., Serra, E. & Spezzano, F. Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums. *Netw. Modeling Anal. Health Inform. Bioinforma.* **9**, 1–11 (2020).
106. Friedenberg, M., Amiri, H., Daumé III, H. & Resnik, P. The umd clpsych 2016 shared task system: text representation for predicting triage of forum posts about mental health. In *Proc. Third Workshop on Computational Linguistics and Clinical Psychology*, pp. 158–161 (2016).
107. Nguyen, T. Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimed. Tools Appl.* **76**, 10653–10676 (2017).
108. Yazdavar, A. H. et al. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proc. 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 1191–1198 (2017).
109. Sarsam, S. M., Al-Samarraie, H., Alzahrani, A. I., Alnumay, W. & Smith, A. P. A lexicon-based approach to detecting suicide-related messages on twitter. *Biomed. Signal Process. Control* **65**, 102355 (2021).
110. Driessens, K., Reutemann, P., Pfahringer, B. & Leschi, C. Using weighted nearest neighbor to benefit from unlabeled data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 60–69 (2006).
111. Zhou, D., Bousquet, O., Lal, T. N., Weston, J. & Schölkopf, B. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pp. 321–328 (2004).
112. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
113. Voulodimos, A., Doulamis, N., Doulamis, A. & Protopapadakis, E. Deep learning for computer vision: a brief review. *Comput. Intell. Neurosci.* **2018** 1–13 (2018).
114. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **13**, 55–75 (2018).
115. Deng, L. & Yu, D. Deep learning: methods and applications. *Found. Trends Signal Process.* **7**, 197–387 (2014).
116. Su, C., Xu, Z., Pathak, J. & Wang, F. Deep learning in mental health outcome research: a scoping review. *Transl. Psychiatry* **10**, 1–26 (2020).
117. Ghannay, S., Favre, B., Esteve, Y. & Camelin, N. Word embedding evaluation and combination. In *Proc. Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 300–305 (2016).
118. Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543 (2014).
119. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. In *Proc. 1st International Conference on Learning Representations (ICLR) Workshops Track*. (2013).
120. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for languageunderstanding. In *Proc. NAACL-HLT*, pp. 4171–4186 (2019).
121. Lan, Z. et al. Albert: a lite bert for self-supervised learning of language representations. In *Proc. 8th International Conference on Learning Representations* (ICLR) (2020).
122. Gaur, M. et al. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, pp. 514–525 (2019).
123. Boukil, S., El Adnani, F., Cherrat, L., El Moutaouakkil, A. E. & Ezziyyani, M. Deep learning algorithm for suicide sentiment prediction. In *International Conference on Advanced Intelligent Systems for Sustainable Development*, pp. 261–272 (2018).
124. Phan, H. T., Tran, V. C., Nguyen, N. T. & Hwang, D. A framework for detecting user's psychological tendencies on twitter based on tweets sentiment analysis. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 357–372 (2020).
125. Wang, Y. -T., Huang, H. -H., Chen, H. -H. & Chen, H. A neural network approach to early risk detection of depression and anorexia on social media text. In *CLEF (Working Notes)* (2018).
126. Trotzek, M., Koitka, S. & Friedrich, C. M. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Trans. Knowl. Data Eng.* **32**, 588–601 (2018).
127. Obeid, J. S. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. *BMC Med. Inform. Decis. Mak.* **19**, 1–9 (2019).
128. Rao, G., Zhang, Y., Zhang, L., Cong, Q. & Feng, Z. Mgl-cnn: a hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access* **8**, 32395–32403 (2020).
129. Lin, H. Detecting stress based on social interactions in social networks. *IEEE Trans. Knowl. Data Eng.* **29**, 1820–1833 (2017).
130. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
131. Cho, K. et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proc. the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734 (2014).
132. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pp. 1310–1318 (2013).
133. Ghosh, S. & Anwar, T. Depression intensity estimation via social media: a deep learning approach. *IEEE Trans. Comput. Soc. Syst.* **8**, 1465–1474 (2021).
134. Uddin, A. H., Bapery, D. & Arif, A. S. M. Depression analysis of bangla social media data using gated recurrent neural network. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1–6 (2019).
135. Yao, H. Detection of suicidality among opioid users on reddit: machine learning–based approach. *J. Med. Internet Res.* **22**, 15293 (2020).
136. Ahmed, U., Mukhiya, S. K., Srivastava, G., Lamo, Y. & Lin, J. C. -W. Attention-based deep entropy active learning using lexical algorithm for mental health treatment. *Front. Psychol.* **12**, 471 (2021).
137. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proc. 3rd International Conference on Learning Representations (ICLR)* (2015).
138. Sekulić, I. & Strube, M. Adapting deep learning methods for mental health prediction on social media. In *Proc. the 5th Workshop on Noisy User-generated Text (W-NUT)*, pp. 322–327 (2019).
139. Sawhney, R., Joshi, H., Gandhi, S. & Shah, R. R. Towards ordinal suicide ideation detection on social media. in: *Proc. 14th ACM International Conference on Web Search and Data Mining*, pp. 22–30 (2021).
140. Rutowski, T. et al. Cross-demographic portability of deep nlp-based depression models. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1052–1057 (2021).
141. Rutowski, T. et al. Depression and anxiety prediction using deep language models and transfer learning. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pp. 1–6 (2020).
142. Ghosh, S., Ekbal, A. & Bhattacharyya, P. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognit. Comput.* **14**, 110–129 (2022).
143. Gui, T. et al. Cooperative multimodal approach to depression detection in twitter. In *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, pp. 110–117 (2019).
144. Gui, T. et al. Depression detection on social media with reinforcement learning. In *China National Conference on Chinese Computational Linguistics*, pp. 613–624 (2019).
145. Wongkoblap, A., Vadillo, M. A. & Curcin, V. Predicting social network users with depression from simulated temporal data. In *IEEE EUROCON 2019-18th International Conference on Smart Technologies*, pp. 1–6 (2019).
146. Wongkoblap, A., Vadillo, M. A. & Curcin, V. Modeling depression symptoms from social network data through multiple instance learning. *AMIA Summits Transl. Sci. Proc.* **2019**, 44 (2019).
147. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008 (2017).
148. Wang, N. et al. Learning models for suicide prediction from social media posts. In *Proc. the Seventh Workshop on Computational Linguistics and Clinical Psychology*, pp. 87–92 (2021).
149. Zhang, T., Schoene, A. M. & Ananiadou, S. Automatic identification of suicide notes with a transformer-based deep learning model. *Internet Interv.* **25**, 100422 (2021).

150. Haque, F., Nur, R. U., Al Jahan, S., Mahmud, Z. & Shah, F. M. A transformer based approach to detect suicidal ideation using pre-trained language models. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5 (2020).
151. Chaurasia, A. et al. Predicting mental health of scholars using contextual word embedding. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 923–930 (2021).
152. Malviya, K., Roy, B. & Saritha, S. A transformers approach to detect depression in social media. In *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, pp. 718–723 (2021).
153. Murarka, A., Radhakrishnan, B. & Ravichandran, S. Detection and classification of mental illnesses on social media using roberta. Preprint *arXiv* https://arxiv.org/abs/2011.11226 (2020).
154. Wang, X. Depression risk prediction for chinese microblogs via deep-learning methods: content analysis. *JMIR Med. Inform.* **8**, 17958 (2020).
155. Abed-Esfahani, P. et al. Transfer learning for depression: early detection and severity prediction from social media postings. In *CLEF (Working Notes)* (2019).
156. Gaur, M. Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs. *PloS ONE* **16**, 0250448 (2021).
157. Tadesse, M. M., Lin, H., Xu, B. & Yang, L. Detection of suicide ideation in social media forums using deep learning. *Algorithms* **13**, 7 (2020).
158. Zhou, S., Zhao, Y., Bian, J., Haynos, A. F. & Zhang, R. Exploring eating disorder topics on twitter: machine learning approach. *JMIR Med. Inform.* **8**, 18273 (2020).
159. Deshpande, S. & Warren, J. Self-harm detection for mental health chatbots. In *Public Health and Informatics*, pp. 48–52. (IOS Press, 2021).
160. Solieman, H. & Pustozerov, E. A. The detection of depression using multimodal models based on text and voice quality features. In *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, pp. 1843–1848 (2021).
161. Sawhney, R., Joshi, H., Gandhi, S. & Shah, R. A time-aware transformer based model for suicide ideation detection on social media. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7685–7697 (2020).
162. Plutchik, R. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pp. 3–33. (Elsevier, 1980).
163. Aragón, M. E., López-Monroy, A. P., González-Gurrola, L. C. & Montes, M. Detecting depression in social media using fine-grained emotions. in: *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1481–1486 (2019).
164. Aragón, M. E., López-Monroy, A. P., González, L. C. & Montes-y-Gómez, M. Attention to emotions: detecting mental disorders in social media. In *International Conference on Text, Speech, and Dialogue*, pp. 231–239 (2020).
165. Lara, J. S., Aragon, M. E., Gonzalez, F. A. & Montes-y-Gomez, M. Deep bag-of-sub-emotions for depression detection in social media. In *Proc. International Conference on Text, Speech, and Dialogue*. pp. 60–72 (2021).
166. Sawhney, R., Joshi, H., Flek, L. & Shah, R. Phase: Learning emotional phase-aware representations for suicide ideation detection on social media. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2415–2428 (2021).
167. Sawhney, R., Joshi, H., Shah, R. & Flek, L. Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2176–2190 (2021).
168. Thorstad, R. & Wolff, P. Predicting future mental illness from social media: a big-data approach. *Behav. Res. Methods* **51**, 1586–1600 (2019).
169. Aladağ, A. E., Muderrisoglu, S., Akbas, N. B., Zahmacioglu, O. & Bingol, H. O. Detecting suicidal ideation on forums: proof-of-concept study. *J. Med. Internet Res.* **20**, 9840 (2018).
170. Desmet, B. & Hoste, V. Online suicide prevention through optimised text classification. *Inf. Sci.* **439**, 61–78 (2018).
171. Cheng, Q., Li, T. M., Kwok, C.-L., Zhu, T. & Yip, P. S. Assessing suicide risk and emotional distress in chinese social media: a text mining and machine learning study. *J. Med. internet Res.* **19**, 243 (2017).
172. Roy, A. A machine learning approach predicts future risk to suicidal ideation from social media data. *NPJ Digital Med.* **3**, 1–12 (2020).
173. Rios, A. & Kavuluru, R. Ordinal convolutional neural networks for predicting rdoc positive valence psychiatric symptom severity scores. *J. Biomed. Inform.* **75**, 85–93 (2017).
174. Losada, D. E., Crestani, F. & Parapar, J. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 340–357 (2019).
175. Losada, D. E. & Crestani, F. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 28–39 (2016).
176. Van Engelen, J. E. & Hoos, H. H. A survey on semi-supervised learning. *Mach. Learn.* **109**, 373–440 (2020).
177. Settles, B. Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances. In *Proc. 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1467–1478 (2011).
178. Maupomé, D. & Meurs, M. -J. Using topic extraction on social media content for the early detection of depression. In *CLEF (Working Notes)* vol. 2125 (2018)
179. Gaur, M. et al. "Let me tell you about your mental health!" contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proc. 27th ACM International Conference on Information and Knowledge Management*, pp. 753–762 (2018).
180. Galiatsatos, D. et al. Classification of the most significant psychological symptoms in mental patients with depression using bayesian network. In *Proc. 16th International Conference on Engineering Applications of Neural Networks (INNS)*, pp. 1–8 (2015).
181. Wang, W. Y., Singh, S. & Li, J. Deep adversarial learning for nlp. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 1–5 (2019).
182. Le-Khac, P. H., Healy, G. & Smeaton, A. F. Contrastive representation learning: a framework and review. *IEEE Access* **8**, 193907–193934 (2020).
183. Li, Y., Tian, X., Liu, T. & Tao, D. Multi-task model and feature joint learning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015).
184. Sharma, A. R. & Kaushik, P. Literature survey of statistical, deep and reinforcement learning in natural language processing. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pp. 350–354 (2017).
185. Ruder, S., Peters, M. E., Swayamdipta, S. & Wolf, T. Transfer learning in natural language processing. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pp. 15–18 (2019).
186. López-Úbeda, P., Plaza-del-Arco, F. M., Díaz-Galiano, M. C. & Martín-Valdivia, M.-T. How successful is transfer learning for detecting anorexia on social media? *Appl. Sci.* **11**, 1838 (2021).
187. Hu, D. An introductory survey on attention mechanisms in nlp problems. In *Proc. SAI Intelligent Systems Conference*, pp. 432–448 (2019).
188. Wang, Z., Zhang, J., Feng, J. & Chen, Z. Knowledge graph and text jointly embedding. In *Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1591–1601 (2014).
189. Sap, M., Shwartz, V., Bosselut, A., Choi, Y. & Roth, D. Commonsense reasoning for natural language processing. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 27–33 (2020).
190. Feng, S. Y. et al. A survey of data augmentation approaches for nlp. In *Proc. Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pp. 968–988 (2021).
191. Lin, C. et al. Sensemood: depression detection on social media. In *Proc. 2020 International Conference on Multimedia Retrieval*, pp. 407–411 (2020).
192. Mann, P., Paes, A. & Matsushima, E. H. See and read: detecting depression symptoms in higher education students using multimodal social media data. In *Proc. International AAAI Conference on Web and Social Media*, vol. 14, pp. 440–451 (2020).
193. Xu, Z., Pérez-Rosas, V. & Mihalcea, R. Inferring social media users' mental health status from multimodal information. In *Proc. 12th Language Resources and Evaluation Conference*, pp. 6292–6299 (2020).
194. Wang, B. et al. Learning to detect bipolar disorder and borderline personality disorder with language and speech in non-clinical interviews. In *Proc. Interspeech 2020*, pp. 437–441 (2020).
195. Rodrigues Makiuchi, M., Warnita, T., Uto, K. & Shinoda, K. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proc. 9th International on Audio/Visual Emotion Challenge and Workshop*, pp. 55–63 (2019).
196. Mittal, A. et al. Multi-modal detection of alzheimer's disease from speech and text. In *Proc. BIOKDD'21* (2021).
197. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why should i trust you?" explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144 (2016).
198. Du, M., Liu, N. & Hu, X. Techniques for interpretable machine learning. *Commun. ACM* **63**, 68–77 (2019).
199. Song, H., You, J., Chung, J. -W. & Park, J. C. Feature attention network: Interpretable depression detection from social media. In *Proc. 32nd Pacific Asia Conference on Language, Information and Computation* (2018).
200. Zogan, H., Razzak, I., Wang, X., Jameel, S. & Xu, G. Explainable depression detection with multi-modalities using a hybrid deep learning model on social media. Preprint at *arXiv* https://arxiv.org/abs/2007.02847 (2020).
201. Castelvecchi, D. Can we open the black box of AI? *Nat. N.* **538**, 20 (2016).
202. Benton, A., Coppersmith, G. & Dredze, M. Ethical research protocols for social media health research. In *Proc. First ACL Workshop on Ethics in Natural Language Processing*, pp. 94–102 (2017).

203. Nicholas, J., Onie, S. & Larsen, M. E. Ethics and privacy in social media research for mental health. *Curr. Psychiatry Rep.* **22**, 1–7 (2020).
204. McKee, R. Ethical issues in using social media for health and health care research. *Health Policy* **110**, 298–301 (2013).
205. Tadisetty, S. & Ghazinour, K. Anonymous prediction of mental illness in social media. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0954–0960 (2021).
206. Doan, S. Extracting health-related causality from twitter messages using natural language processing. *BMC Med. Inform. Decis. Mak.* **19**, 71–77 (2019).
207. Hutto, C. & Gilbert, E. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In *Proc. International AAAI Conference on Web and Social Media*, vol. 8 (2014).
208. Cambria, E., Poria, S., Hazarika, D. & Kwok, K. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proc. AAAI Conference on Artificial Intelligence*, vol. 32 (2018).
209. Nielsen, F. Å. A new anew: evaluation of a word list for sentiment analysis in microblogs. In *Proc. CEUR Workshop Proceedings*, vol. 718, pp. 93–98 (2011).
210. Wang, X. et al. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 201–213 (2013).
211. Leiva, V. & Freire, A. Towards suicide prevention: early detection of depression on social media. In *International Conference on Internet Science*, pp. 428–436 (2017).
212. Stephen, J. J. & Prabu, P. Detecting the magnitude of depression in twitter users using sentiment analysis. *Int. J. Electr. Comput. Eng.* **9**, 3247 (2019).
213. Mohammad, S. M. & Turney, P. D. Nrc emotion lexicon. National Research Council, Canada **2** (2013).
214. Zhou, T. H., Hu, G. L. & Wang, L. Psychological disorder identifying method based on emotion perception over social networks. *Int. J. Environ. Res. Public Health* **16**, 953 (2019).
215. Saloun, P., Ondrejka, A., Malčík, M. & Zelinka, I. Personality disorders identification in written texts. In *AETA 2015: Recent Advances in Electrical Engineering and Related Sciences*, pp. 143–154 (Springer, 2016).
216. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
217. Dumais, S. T. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.* **38**, 188–230 (2004).
218. Xu, W., Liu, X. & Gong, Y. Document clustering based on non-negative matrix factorization. In *Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 267–273 (2003).
219. Desmet, B., Jacobs, G. & Hoste, V. Mental distress detection and triage in forum posts: the lt3 clpsych 2016 shared task system. In *Proc. Third Workshop on Computational Linguistics and Clinical Psychology*, pp. 148–152 (2016).
220. Tausczik, Y. R. & Pennebaker, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**, 24–54 (2010).
221. Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G. & Rosa, T. C. Sentihealth-cancer: a sentiment analysis tool to help detecting mood of patients in online social networks. *Int. J. Med. Inform.* **85**, 80–95 (2016).
222. Yoo, M., Lee, S. & Ha, T. Semantic network analysis for understanding user experiences of bipolar and depressive disorders on reddit. *Inf. Process. Manag.* **56**, 1565–1575 (2019).
223. Ricard, B. J., Marsch, L. A., Crosier, B. & Hassanpour, S. Exploring the utility of community-generated social media content for detecting depression: an analytical study on instagram. *J. Med. Internet Res.* **20**, 11817 (2018).
224. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119 (2013).
225. Hemmatirad, K., Bagherzadeh, H., Fazl-Ersi, E. & Vahedian, A. Detection of mental illness risk on social media through multi-level svms. In *2020 8th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS)*, pp. 116–120 (2020).
226. Bandyopadhyay, A., Achilles, L., Mandl, T., Mitra, M. & Saha, S. K. Identification of depression strength for users of online platforms: a comparison of text retrieval approaches. In *Proc. CEUR Workshop Proceedings*, vol. 2454, pp. 331–342 (2019).
227. Zhong, Q. -Y. 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.* **18**, 1–11 (2018).
228. Huang, Y., Liu, X. & Zhu, T. Suicidal ideation detection via social media analytics. In *International Conference on Human Centered Computing*, pp. 166–174 (2019).
229. Lv, M., Li, A., Liu, T. & Zhu, T. Creating a chinese suicide dictionary for identifying suicide risk on social media. *PeerJ* **3**, 1455 (2015).
230. Nguyen, T., Phung, D., Adams, B. & Venkatesh, S. Prediction of age, sentiment, and connectivity from social media text. In *International Conference on Web Information Systems Engineering*, pp. 227–240 (2011).
231. Peng, Z., Hu, Q. & Dang, J. Multi-kernel svm based depression recognition using social media data. *Int. J. Mach. Learn. Cybern.* **10**, 43–57 (2019).
232. Wu, M. Y., Shen, C.-Y., Wang, E. T. & Chen, A. L. A deep architecture for depression detection using posting, behavior, and living environment data. *J. Intell. Inf. Syst.* **54**, 225–244 (2020).
233. Zogan, H., Wang, X., Jameel, S. & Xu, G. Depression detection with multi-modalities using a hybrid deep learning model on social media. Preprint at *arXiv* https://arxiv.org/abs/2007.02847 (2020).
234. Yao, X., Yu, G., Tang, J. & Zhang, J. Extracting depressive symptoms and their associations from an online depression community. *Comput. Hum. Behav.* **120**, 106734 (2021).
235. Dinkel, H., Wu, M. & Yu, K. Text-based depression detection on sparse data. Preprint at *arXiv* https://arxiv.org/abs/1904.05154 (2019).
236. Zhou, Y., Glenn, C. & Luo, J. Understanding and predicting multiple risky behaviors from social media. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence* (2017).
237. Wang, Y., Wang, Z., Li, C., Zhang, Y. & Wang, H. A multitask deep learning approach for user depression detection on sina weibo. Preprint at *arXiv* https://arxiv.org/abs/2008.11708 (2020).
238. Aragon, M. E., Lopez-Monroy, A. P., Gonzalez-Gurrola, L. -C. G. & Montes, M. Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression. *IEEE Trans. Affect. Comput*. (2021).
239. Li, N., Zhang, H. & Feng, L. Incorporating forthcoming events and personality traits in social media based stress prediction. *IEEE Trans. Affect. Comput.* (2021).

ACKNOWLEDGEMENTS

This research was partially funded by the Alan Turing Institute and the H2020 EPHOR project, grant agreement No. 874703.

AUTHOR CONTRIBUTIONS

T.Z. conducted the review, prepared figures, and wrote the initial draft. A.M.S., S.J., and S.A. revised the paper. S.A. supervised the project. All authors reviewed the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00589-7.

Correspondence and requests for materials should be addressed to Sophia Ananiadou.