Annotation Practices in Machine Learning Research On Depression

Aleksandra Andrasz
Supervisor(s): Cynthia C. S. Liem, Andrew M. Demetriou
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Aleksandra Andrasz
Contact: a.andrasz@student.tudelft.nl
Final project course: CSE3000 Research Project
Thesis committee: Cynthia C. S. Liem, Andrew M. Demetriou, Frank Broz

Abstract

Depression diagnosis and treatment remain difficult tasks that could be improved with machine learning models. But those automatic systems should be reliable to apply in clinical psychology settings. Performing predictions in this field is most commonly done using supervised learning models, which rely on well-established annotations. Therefore this paper examines annotation practices in research surrounding depression and provides a perspective on the quality of established methods. Firstly, 80 papers were surveyed in terms of reported annotation practices. Then the results were collected and analyzed. The findings suggest that papers from the Computer Science domain would benefit from the utilization of expert knowledge and better practices of human verification. While papers across all domains missed important information on the annotation process and rarely (20% of papers) provided input data.

1 Introduction

It is estimated that 4,4% of adults worldwide suffered from depression in 2015 [1], which puts significant strain on the healthcare systems. [2] approximated that in 2020 in the US economic burden of Major Depressive Disorder (MDD) was 326.2 billion dollars. With high treatment and diagnosis costs, many patients do not receive appropriate care. Difficulty in performing reliable diagnosis and treatment being effective in 30-50% of times[3] additionally prolongs suffering for depressed people. An effort in psychology and psychiatry research has been made to increase the efficiency of therapy but depression remains difficult and expensive to treat. [3].

In clinical and research practices, the major classification systems Diagnostic and Statistical Manual of Mental Disorders (DSM) [4] and International Classification of Diseases (ICD) [5] serve as a basis for the definition of depression [6]. A diagnosis can be then performed by comparing patients' symptoms to those defined in DSM or ICD. A structured way of measuring those symptoms can be done by administering a depression-rating scale. Such an instrument measures depression through a set of questions, each question is rated and the score determines depression severity. Clinician-rated instruments such as Montgomery Åsberg Depression Rating Scale (MADRS) [7] are administered by trained clinicians through interviewing the patient. This approach can be costly because it requires additional training for the expert and prolongs diagnosis. However, self-reporting scales, such as Beck Depression Inventory (BDI) [8], which can be completed by the patient, do not provide complete information on patients' health [2]. This leaves the diagnosis process complicated and difficult to individualize [3].

Machine learning provides an opportunity to make diagnosis and treatment of depression more affordable and available [9]. Automatic systems could provide useful insight in the early stages of treatment, allowing for a more personalized approach. It can be achieved by utilizing machine learning systems trained on patient data. It is essential that those systems are reliable because well-performed and transparent prediction becomes a bottleneck in applying the research in practice [3].

Supervised learning models are most widely used in depression research [3], those models are trained and verified against established reference - also called "ground truth". Often humans are asked to label the input data beforehand to determine the expected output of the trained model. Later such ground truth is provided as a training and validation set for the model, to train it and access its efficacy. Therefore the quality of prediction, as well as the accuracy score, is dependent on well-collected and labelled data.

This research examines the quality of ground truth in Machine Learning research surrounding depression, focusing on annotation practices. Annotation practice means the process of collecting ground truth labels for training and test data sets. Relatively few research papers examine data annotation practices which constitute model reliability. Geiger et. al performed a survey on papers available on *arxiv.org* and showed few papers exhibit good annotation practices [10]. Another research, focusing on mental health prediction in social media [11], brought attention to the lack of established annotation practices in the field.

This paper answers the following question: *What are the data collection and reporting practices of annotations within Machine Learning Research surrounding depression?* To that end, the paper first examines the most cited papers performing machine learning research to predict Major Depressive Disorder (MDD), its symptoms and treatment outcomes. Then the results are collected and analyzed. For each paper, sub-questions are answered and the results are summarized and interpreted. The main research question can be divided into the following sub-questions:

- *What are the sources of used labels?*
- *What is the quality of reporting practices of human annotation process?*
- *What is reported about ground truth data?*
- *How does the quality vary between different research domains?*

Therefore this research provides an overview of annotation practice in depression research and a perspective on possible improvements. The results of the analysis suggest that Medicine and Psychiatry papers differed in their reporting practice, compared to Computer Science and Other domains. The most substantial difference was found to be in reporting of formal instructions and training. Additionally, the author noted the issues that could be improved across all domains, those concerned availability of data and transparency of annotation practices.

The text is structured in the following way. Section 2 describes the selection, collection of papers and their analysis. Section 3 provides described collected results. In Section 4, the author provides their perspective on obtained results. Section 5 deals with the ethical implications and reproducibility of the research. Lastly, Section 6 concludes the report and provides future directions for reaserch.

2 Data and Methods

This section described the used methods to answer posed research question. The Subsection 2.1 describes search queries and inclusion and exclusion criteria. Then Subsection 2.2 details the subsequent examination of chosen papers by listing all questions.

2.1 Data Collection

Papers were collected from the Scopus[1] database from the 2nd of May until the 23rd of May 2023. Search queries were determined based on the research question. The final search query is present in Table 1. The search query itself was divided into several search blocks, to utilize the Scopus interface that enables dynamic adjustment of the search query.

The first part identified depression research. Included keywords included depression synonyms (“major depressive disorder”), and symptoms commonly associated with the disorder (“low mood”, “suicidality”). The next part limited the scope to the machine learning domain, keywords such as “deep learning”, “statistical classifier” along with “machine learning” were added. Papers referred to “machine learners”, therefore keywords were augmented with “*” (ex. “machine learn*”). Since the research focuses on depression as a mental health disorder, the keywords: “human”, “patient”, and “people” were incorporated, to exclude any other meanings of the word. For example in image processing research (background depression). All parts were connected using the “AND” operator. Full search query is visible in Table 1.

The search query had to be adjusted several times because it became apparent that certain papers were neglected while others should be excluded. For example, after screening the literature it became necessary to add support vector machines (SVM) as a keyword. SVMs were often deployed to predict depression from EEG scans. The keyword: “st depression” was appended with the “AND NOT” operator to exclude papers that were not eligible for the review. The last search was performed on the 23rd of May 2023.

Finally, the papers were examined in order of most cited according to Scopus. The author read the paper’s abstract and determined the topic. The goal was to include papers that applied machine learning to predict depression, symptoms or treatment. Document type “review” was excluded since papers performing prediction were the focus. A limited time frame, from 2013-2023, featured papers that exhibit current trends. While allowing to take into account a trend to perform mental health predictions on social media which was present since around 2013 (highly cited [12]). The final overview of inclusion/exclusion criteria is present in Table 2.

[1] https://www.scopus.com

Table 1: Table with keywords from performed search.

Keywords	
Depression related	depression, “depressive disorder*”, “depress* symptom*”, “low mood*”, “mood disorder*”, “postpartum depress*”, suicidality, suicidal, suicide, “self harm*”, “social withdrawal”
Machine Learning related	“Bayes classifier*”, “automat* classification”, “statistical classifier*”, “probabilistic classifier*”, “supervised learn*”, “machine learn*”, “deep learn*”, “artificial intelligence”, “machine model*”, “supervised model*”, “Support Vector Machine”, “convolutional neural network*” , “random forest*”, “natural language processing”, “learning algorithm*”, “probabilistic model*”, “predicti* framework*”, “statistical framework*”
Human related	individual*, user*, mental*, human*, people, person*, patient*, wom?n, m?n
Excluding irrelevant papers (AND NOT)	“subthreshold depression”, “ST depression”

Table 2: Table with Inclusion and Exclusion criteria.

Inclusion and Exclusion criteria		
Category	Included papers	Excluded papers
Year	2013 - 2023	Before 2013
Paper type	Other	Review or survey
Topic	Paper applies machine learning models to predict depression, symptoms or treatment	Other
Peer-review	Peer-reviewed	Paper is not peer-reviewed
Language	English	non-English

2.2 Data Analysis

Papers were examined by the author to review collection methods of ground truth data. This section lists all of the questions that were used for the analysis process. Most of them were sourced from Geiger et. al [10], but compared to [10] the author searched for available information outside of the original paper. Answers to each question were recorded for all papers included in the study. The questions estimate the quality of the annotation process the reasoning behind each of them is provided. Each sub-subsection is connected to one of the research sub-questions.

The author used an Excel sheet to store the results and analyze them[2]. The answer to each question was recorded in a separate column and for some questions, there was an additional column for unstructured notes. Those notes were incorporated into the results at the end. When possible, the author created labels corresponding to possible answers for each question. This was done to minimize variance between the labelling of each paper. When the information was not provided label ‘unsure’ was used.

2.2.1 Papers’ categories

Machine learning research on depression is a broad field that researchers from various domains try to tackle (Medicine, Computer Science, Psychology, Neuroscience, etc.). Coming from the Computer Science perspective most interest was in how computer scientists try to take the problem and how it compares to other domains. Moreover, the field of Medicine might employ more rigorous methods of data annotation while tackling research on depression. Therefore the papers are divided into separate domains.

The author used information available on Scopus. The platform provides information on the research facility of the authors often with their department name. If the information was not specified the author used Google search engine to determine it. The information was chosen as a basis for the

[2] https://github.com/aleksandra-and/CSE3000-Research-Project-2022-23-Q4-

category because it was easily available and reflected the domain of the authors. Results could be analysed per category.

2.2.2 Questions on types of labels

The first three questions examine external sources used by the reviewed papers. Those questions also determine the inclusion of additional studies in the review: **1)** ***was the work an original task?***. This question verifies if examined paper trained an original machine-learning model. Otherwise, a paper detailing the external pre-trained model has to be reviewed. **2)** ***did the paper's authors created (original) annotations/data set?*** **3)** ***did the paper's authors use existing (external) annotations/data set?*** Those two questions verify if the authors used existing data set or labels. If a paper used more than one external data set all papers detailing the creation of such data sets are included in the review. Next questions were answered separately for each data set that matched the research.

The next question identified papers that appointed human annotators to create labels: **4)** ***did they use human annotations?***. This question filters papers that tasked humans with assigning labels to multiple data points. Another way of sourcing labels utilizing human input would be using self-reports on depression. Self-reports were not considered human annotation in the paper. The author chose this definition because subsequent questions (Subsubsection 2.2.3) assess the quality of labels created by separately appointed annotators that labelled multiple items.

Many papers only briefly mention the use of human annotations or declare it through psychiatric instruments. As an example Montgomery–Åsberg Depression Rating Scale (MADRS) [7], a scale that determines the severity of depression, should be administered by a clinician. Therefore if the authors report the use of such an instrument, humans should have annotated the data. Though it was not always explicitly mentioned, the author decided to treat it as human annotations.

The next question explores annotations not based on human input: but instead based on machine learning models, psychometric scales and other sources: **5)** ***did the paper's authors use non-human annotations?***. This question examines if the data set labels are based on other than human input. Self-reports of depression are considered to be in this category. **6)** ***what kind of non-human annotations did the paper's authors use?*** further details the non-human input. Unstructured answers to the 6th question were combined after the data analysis process. Those two questions explored depression self-reporting scales scores and filtered social media posts used as ground truth labels.

2.2.3 Questions on quality of human annotations

The next questions concern the quality of human annotations. First examines **7)** ***who were the annotators?***. The answer was recorded according to what was stated in the paper. At the end of the data analysis process the author combined the answers to one of the options: 'experts', 'non-experts', 'experts and non-experts' and 'unsure'. The non-expert label refers to people who do not have professional expertise or training in the depression domain, while experts would be trained clinicians. Depression is a complex problem and the costs of a mistake are high, therefore knowledge of a person making the decision should be reported.

Furthermore, to check the number of annotators that were appointed in the creation of the data set question: **8)** ***did they report the number of annotators?*** was asked. The paper could only obtain a label yes if the exact number was reported. It might be more appropriate to appoint a larger number of annotators for bigger data sets.

Defining depression is a complicated task, such definitions are usually part of extensive manuals [4][5]. Therefore it might be difficult to achieve consistent and reliable labelling. A way of ensuring consistency among multiple annotators is by providing formal instructions. Next two questions focused on those aspects: **9)** ***did the paper's authors report providing formal instructions for the annotators?***. Those had to be formal definitions or established manuals. A short description (one sentence) or examples of annotated data was not enough. These criteria were chosen because depression is a complex disease and short instructions introduce variance in interpretation. The approach has limitations: an experienced psychiatrist could diagnose depression without formal instructions.

Another way of ensuring label quality is training for annotators: **10)** ***did the paper's authors report training the annotators?***. The paper was annotated with 'yes' when training was mentioned or all the annotators were experts. The assumption is that even though experts can benefit from additional training it is less crucial than training for non-export annotators.

To further improve the quality of labels multiple annotators could label the same data items: **11)** ***was there multiple-annotator overlap?***. Paper received 'yes' to this question if it reported annotators labelling the same set of items. Inter-annotator agreement was recorded with the next question: **12)** ***did the paper's authors report inter-annotator agreement?***. Inter-annotator agreement metrics are a way to report overlap between answers of different annotators. Question: **13)** ***did the paper's authors report other metrics of label quality?*** addressed a case when authors used a different metric to measure label quality.

2.2.4 Questions on data quality

The availability of ground truth data contributes to the reliability of the machine learning system and the reproducibility of the research. The author answered the questions: **14)** ***did the paper's authors link to the data set?***, and **15)** ***would it be possible to retrieve the data?***. With question 14 the author searched for links to used data sets or their description. While question 15 examined if the data was attainable (usable links, external repository, available on request, etc.). Those two questions give an overview of the data availability in examined research.

Determining prediction around mental health is difficult and relies on individualizing diagnosis and care process [3]. Therefore for generalizable results, it is important to have a representative data sample. The author briefly examines what information is available on data used by review papers: **16)** ***what information was reported about demographics?***. An example answers to this question could be: 'no information' or provided: 'age, gender, race'. This is a limited approach but it can suggest possible future extensions of the study.

3 Results

In this section, the author characterizes gathered papers and answers questions detailed in Subsection 2.2. Subsection 3.1 describes the resulting study selection. Next Subsection 3.2 described what label sources were used. In Subsection 3.4 the author focuses on the characteristics of human-annotations. Lastly, Subsection 3.5 details the information the author gathered on data used by reviewed papers.

3.1 Study selection

The final Scopus query matched 4909 papers. The author read abstracts until collecting 80 eligible papers. 121 papers' abstracts were examined in order of most cited according to Scopus. Included papers had from 84 to 1004 citations. A diagram detailing the process is visible in Figure 1. The final set of included papers is visible in Appendix A and used Excel sheet[3].

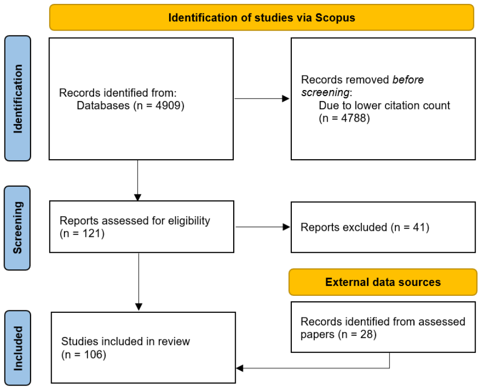
Identification of studies via Scopus
Identification
Records identified from:
Databases (n = 4909)
Records removed *before screening*:
Due to lower citation count
(n = 4788)
Screening
Reports assessed for eligibility
(n = 121)
Reports excluded (n = 41)
External data sources
Records identified from assessed papers (n = 28)
Included
Studies included in review
(n = 106)

Figure 1: Flow chart of included studies, based on PRISMA guidelines [13].

[3] https://github.com/aleksandra-and/CSE3000-Research-Project-2022-23-Q4-

Additionally, 28 papers were added due to insufficient information in the original 80-paper set. Examples of "external data source" are [14], [15] and [16]. [17] determines the ground truth, the mental health of social media users, using a model proposed by [14]. While [18] created a model predicting depression from Twitter using data collected by [15] and [16].

The resulting 104 studies were categorised based on the first author's affiliation. There are 3 categories: Computer Science, Medicine and Psychiatry, and Other. The author decided on this grouping because it gives an insight into Computer Science practices. Medicine and Psychiatry papers exhibited more extensive reporting of the results. The 'Other' category comprises papers from Linguistics, Psychology and other domains. The number of papers for each category is visible in Table 3.

Table 3: Reviewed papers with corresponding categories, based on the first author's affiliation.

Categories of reviewed papers	
Computer Science	45
Medicine and Psychiatry	26
Other	33

3.2 Sources of Annotations

In this section, sources of annotations are categorized. Firstly the number of papers that used original annotations is reported. Then the section details the number of non-human annotations that were encountered, and lastly, papers using human annotations.

3.2.1 Use of external sources

2 out of all papers used pre-trained models instead of creating their own (Question 1). 73 papers described collecting annotations themselves (Question 2). Additionally, 39 papers used external data sets (Question 3). The results can be seen in Table 4 and 5. 11 papers used both external data sets and created data sets themselves. When a paper used multiple data sets to make a prediction, answers to the next questions were conflicting. Therefore, each data set was separately considered and reported. The author reported the next questions for papers detailing only one data set (95 papers).

Table 4: 2) Papers creating data set/annotations

Label	Count
yes	73
no	26
unsure	7

Table 5: 3) Papers using existing data set/annotations

Label	Count
yes	39
no	58
unsure	9

3.2.2 Human vs Non-human annotations

According to the definition in Subsection 2.2, 62 papers appointed separate annotators to label multiple data items and 19 did not (Question 4). For some papers, it was not possible to determine whether they used labels created by people, even after looking through external data sources. Those papers received the label 'unsure', there are 14 papers with this label.

Researchers from the Medicine and Psychiatry domain report using human input (80,6%) more often than other domains. Computer Science and Other domains' papers seem to appoint human annotators less in their research. Human annotators: Computer Science - 64,3%, Other domains - 54,5% of papers.

Many papers declared other ways than human input to source the labels. For example [12] surveys people from Mechanical Turk [4] on their history of depression using a depression-rating scale. As described in Subsection 2.2, this type of label was not considered human annotation. 19 did not use human annotations at all (for example [12]) and 46 both non-human and human input. Non-human input could be a depression-rating scale, machine learning model, medical report or social media text, visible in Table 6. Distribution of use of different methods is similar among domains (Figure 3).

[4] https://www.mturk.com/

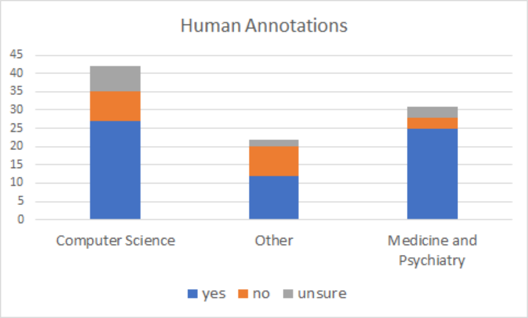

Figure 2: Distribution of examined papers that made use of human annotations.

Table 6: 5) Papers using non-human annotations.

Label	Count
scale	36
model	3
medical report	7
social media text	19
Total	65

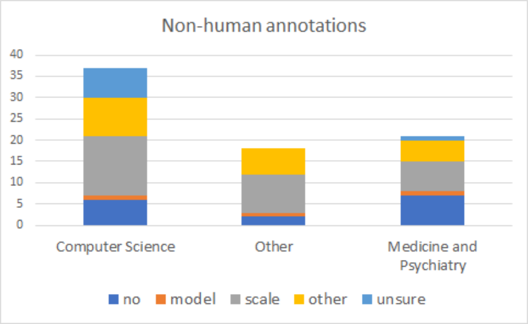

Figure 3: Distribution of examined papers that made use of non-human annotations.

3.3 Non-Human Annotations Quality

This subsection further details non-human sources of annotations. Those are all ways of obtaining annotations that were not considered 'human annotations' in this paper. First Sub-subsection 3.3.1 describes depression-rating scales used in examined research. Then Subsubsection 3.3.2 details other encountered methods.

3.3.1 Depression-rating scales

In depression research (not machine learning related) it is common to use depression-rating scales to determine the severity of Major Depressive Disorder (MDD) in patients [2][19]. Many scales like those were created in 1960s and since then were updated. But the exact version of the used scale was not always available, which brings the question: was the recent version used or the older one, which validity was questioned.

The most common distinction between those scales is a clinician(or observer)-rated and self-reporting scales. As the name suggests one is performed by a trained clinician and the other is completed by the patients themselves. Reviewed papers used both self-reports as well as clinician-rated scales to determine ground truth labels.

The most used clinician-rated scales were the Hamilton Rating Scale of Depression [20] and Structured Clinical Interview for DSM-IV [21], all other common scales are listed in Table 7. Those scales were treated as instructions for the annotators, which in the case of both of those scales should also be trained in the use of the instrument.

Table 7: Papers that used clinician-reported scale to determine labels.

Scale	Count
Hamilton Rating Scale of Depression [20]	15
DSM-IV (Structured Clinical Interview) [21]	9
CIDI [22]	4
MINI [23]	4
MADRS [7]	3

36 papers used self-reporting scales to determine the annotations and 13 papers used it as the sole basis for the prediction. Among the used scales were the Center for Epidemiologic Studies Depression (CES-D) scale [24] and Beck Depression Inventory (BDI) [8], the two most popular self-reporting scales among examined papers. Both of those checks for depression symptoms in the past weeks (one or two, depending on the version) using 20 and 21 questions respectively.

Table 8: Papers that used self-reporting scales. The second column lists the number of papers that used only that scale to determine labels and the last all papers that used that scale.

Scale	Only scale	All papers
BDI or BDI-II [8]	3	15
CES-D [24]	5	5
PHQ-9 [25]	1	5
DASS or DASS-21 [26]	2	3

3.3.2 Other Annotation Methods

Scales were not the only way of determining ground truth labels. Papers used machine learning models, medical reports, and text sourced from social media - Table 6. Medical reports would be previously determined diagnoses, it was not considered human annotation for papers that did not specify how the medical record was established. 4 papers did not declare the use of human annotations in medical reports.

Text sourced from social media would be for example filtered tweets (for example [27]) or Reddit[5] posts taken from mental health-related communities. Sourcing text from social media usually meant that authors used publicly available APIs of chosen platforms (Reddit, Twitter, Facebook, etc). Then performed some form of filtering, for example choosing posts containing “I was diagnosed with depression” text [27][28]. And lastly performed text analysis and model training to obtain a prediction. 5 papers used exclusively this procedure to create the labels. In some cases, the studies also appointed annotators to check if the diagnosis seemed “genuine” like in [28].

3.4 Human Annotations quality

This details ways the author quantified the quality of human-created labels: it reports who were the annotators and how many were appointed, if those people received instructions or training and whether researchers reported label quality measures.

3.4.1 Annotators

The first question examined who was appointed to annotate the data set (Question7). In many cases, papers utilized expert knowledge, this means that trained professionals were asked to determine appropriate labels. Among ‘expert’ annotators were mostly psychiatrists or clinicians. In 32 papers annotators were only experts and in 4 more papers appointed both experts and non-experts. Overall 14 papers relied solely on non-expert labelling. No paper reported using crowd-sourcing as a way to label data. Lastly, 14 papers did not provide information on who annotated the data. The results are visible in Table 9.

[5] https://www.reddit.com/

Table 9: **7)** Types of annotators.

Label	Count
experts	32
experts and non-experts	4
non-experts	14
unsure	14

Table 10: 8) Papers that reported the exact number of annotators.

Label	Count
yes	24
no	40

The next question examines how many papers reported the exact number of annotators (Question 8). 24 papers reported how many annotators were appointed to label the data, from that 20 papers reported assigning less than 5 annotators and the remaining 4 reported more than 20. 40 papers did not provide an exact number of annotators.

Medicine and Psychiatry researchers appointed experts to label data the most often, 67%, compared to 41% for Computer Science and 33% for Other domains. Similarly, papers from Medicine and Psychiatry domains reported the exact number of annotators 72% of the time compared to 55% for Computer Science and 58% for Other domains. Figure 4 and 5 present the distributions.

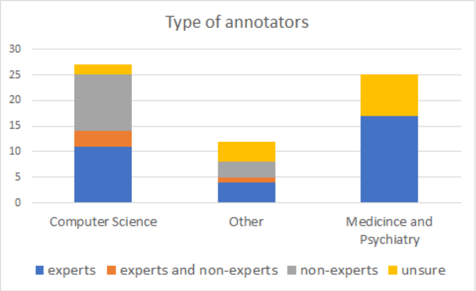

Figure 4: Distribution of types of annotators over the domains.

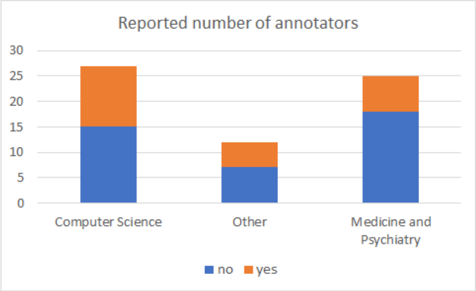

Figure 5: Distribution of the papers that reported the exact number of annotators.

3.4.2 Instructions and Training

37 papers provided formal instruction to the annotators (Question 9) as visible in Table 11. As explained in Section 2.2, one-sentence instructions or examples of annotated items did not constitute 'formal instructions'. An example with shallow instructions is [28], the data set the authors are using is annotated based on whether the self-reported diagnosis seemed genuine. This is a subjective measure which can lead to variance in the results.

34 papers performed training for the annotators (Question 10). When the data set's annotators were experts corresponding paper was assigned a label 'yes' for annotator training. The author encountered papers that performed additional training on expert annotators, for example in [29] "7-day study-specific training program" was performed on expert staff. The resulting number of papers is visible in Table 12.

Table 11: 9) Papers that gave formal instructions for the annotators.

Label	Count
yes	37
no	10
unsure	17

Table 12: 10) Papers that provided training for the annotators.

Label	Count
yes	34
no	4
unsure	26

Researchers from the Medicine and Psychiatry domain report performing training (72%) and giving formal instructions to the annotators (92%) more often than other domains. Reported training might be due to a high number of expert annotators (details Section 2.2). Formal instructions were not directly connected to expert knowledge. Distributions are visible in Table 6 and 7. Computer Science and Other domains papers seem to exhibit lower attention to reporting/performing training and giving instruction to the annotators. Given instructions: Computer Science - 33%, Other domains - 42% of papers. Performed training: Computer Science - 44%, Other domains - 33%.

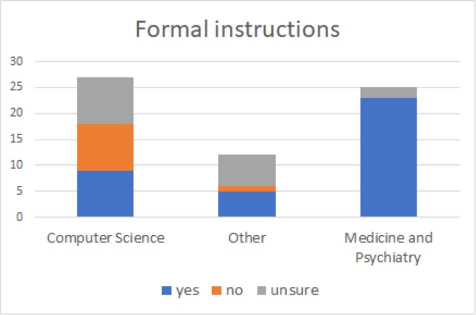

Figure 6: Distribution of papers that reported giving formal instructions to the annotators.

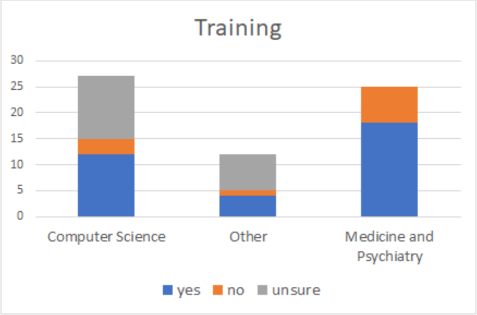

Figure 7: Distribution of papers that reported performing training for the annotators.

3.4.3 Annotator overlap and Label Quality

The next questions review how many papers reported multiple annotators labelling the same item (Question 11) and metrics that were used to assess label quality (Questions 12 and 13). 21 papers had annotators labelling the same set of items, 1 paper reported one annotator and 42 papers did not provide an answer to this question (Question 11). 12 of the papers that reported multiple annotator-overlap also reported the inter-annotator agreement (Table 13). Additionally, 18 papers used other metrics to measure label quality (Table 14), this could be for example using an additional depression-rating scale.

Table 13: 12) Papers calculating inter-annotator agreement

Label	Count
yes	12
no	3
unsure	6

Table 14: 13) Papers using other metrics of label quality.

Label	Count
yes	18
no	46

Interestingly the Computer Science papers reported multiple annotator overlap more often compared to other domains, 44%. Out of those papers, 83% reported inter-annotator agreement (Figure 8). While papers from Medicine and Psychiatry as well as Other domains reported multiple annotator overlap around 25% and inter-annotator agreement 50% and 67% respectively. Medicine and Psychiatry papers used other methods of label quality (44% of papers) while Computer Science (15%) and Other domains (25%) less often did that (Figure 9).

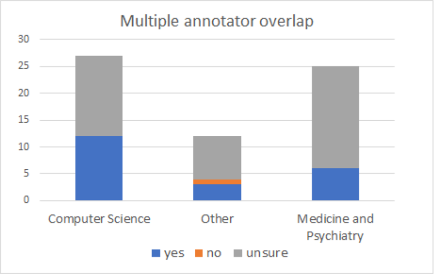

Figure 8: Distribution of papers that reported multiple annotators labelling the same items.

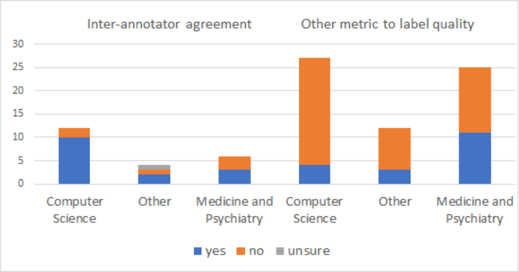

Figure 9: Distribution of papers reporting inter-annotator agreement (on the left) and other metrics of label quality (on the right).

3.5 Information on Input Data

The next questions (Questions 14 and 15) examined whether data used for training and testing was available (Table 15 and 16). When the research provided links to the data, in some cases, the author was able to obtain the data, but other times data was available on request for further research or the links were unusable. While in the case of data available on request, the author noted down that data would be available when the links were unusable label 'no' was chosen. The label 'possibly' (from Table 16) was used in instances when the author believed it was feasible to recreate the data collection process to obtain the same results. For example, when authors used publicly available API and provided dates and filters they used [30].

Table 15: 14) Papers with a link to the data set

Label	Count
yes	31
no	64

Table 16: 15) Papers with available ground truth data

Label	Count
yes	17
possibly	5
no	63

The next questions examined if papers took into account demographic features correlating with depression. Descriptions of the demographics of subjects used in the research (Table 17) were noted. The label 'age, gender and other' refers to a case when a paper report age, gender and other (ex. race, education, income, etc.) information. An example of a paper that reported substantial information on the demographics is: [31], this research describes the distribution of gender, race, marital status, employment and more. On the other hand research such as [32] does not provide any consideration on that matter.

Table 17: 16) Papers with information on demographics of data subjects.

Label	Count
no	52
age	2
gender	4
age, gender	13
age, gender and other	22
gender, region	1
ethnicity, education, income	1

For Medicine and Psychiatry papers the author considered the data to be available for 25% of papers compared to around 14% for the other domains (Figure 10). Therefore no domain has a considerably higher availability of ground truth data. However, 81% papers from Medicine and Psychiatry domain reported demographic information about the data subjects. Some papers from this domain also provided substantial information. Papers labelled with 'age, gender and other' constituted 42% of papers from Medicine and Psychiatry, 10% of Computer Science, and 23% of Other domains (Figure 11).

4 Discussion

This section summarises results and provides possible reasons for observed trends. After collecting a set of papers, each paper was analysed using questions on annotations methods. Papers were categorised into Computer Science, Medicine and Psychiatry, and Other domains. The differences between different domains are further explored in this section. The author also provides an overview of trends visible across domains.

Medicine and Psychiatry papers appointed human annotators more frequently compared to other domains. This is a more costly approach because it requires additional time and resources required for human input. But the verification improves the quality of annotations. For example, the research suggests [19] that when examining optimal treatment options interview by a clinician should be employed along with the use of a self-reporting scale. Therefore using only self-reports might be insufficient to predict depression or treatment. The second most popular method, using filtered

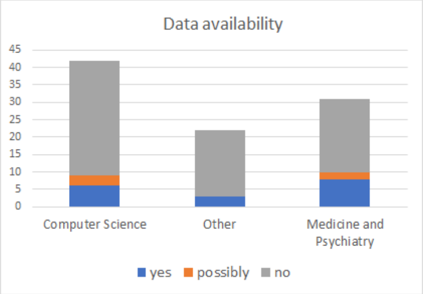

Figure 10: Distribution of papers that had ground truth data available.

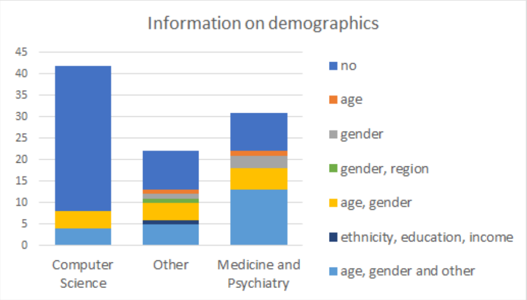

Figure 11: Distribution of papers with demographic information of people who were part of the study.

social media texts, benefits from human verification. As the researchers pointed out [28] some social-media texts could be satirical and thus not reflect real mental health issues. Therefore when using non-human sources, additional human checks would be appropriate.

Similarly, Medicine and Psychiatry papers more often appointed experts to perform the annotation process and provided instructions (72%) and training for the annotators (92%). Expert knowledge can be important when predicting depression and treatment. Generally, the disorder is treated by trained psychiatrists and psychologists, because it is a complex problem and the costs of a mistake are high. For the same reason additional training and formal instructions for the annotators are valuable. An additional benefit of those is decreasing variance between different annotators. So both expert knowledge and training and instructions are beneficial.

However, Computer Science and Other domains appointed multiple annotators to label the same items and reported inter-annotator agreement more often. There are two possible reasons that the author considered. Clinician-rated scales used as formal instruction for the annotators should guarantee high inter-annotator reliability. Therefore some research might choose to rely solely on a clinician-rated scale. Secondly, having multiple expert annotators perform clinical interventions is costly. So the researcher might opt for the use of other methods, such as administering multiple depression-rating scales. Nevertheless, the author's opinion is that multiple annotator overlap and inter-annotator agreement are important methods of increasing label quality.

The area where all papers were consistently lacking was data availability and description. This is a difficult aspect of dealing with data that surrounds depression. For the data to be publicly available it should be anonymised. Reporting the demographics of data was performed by half of the examined papers. This is important because the inclusion of different social groups determines model generalizability across different domains. Depression is also not uniformly distributed in all social groups [1]. The Medicine and Psychiatry domains had slightly higher percentages of papers with demographic descriptions.

Lastly, many reviewed papers lacked information that would allow the author to answer all questions. Among the least provided information was whether multiple annotators labelled the same item and if there was training or formal instructions for the annotators. But for each aspect reviewed in this research, several papers did not specify the answer. There are different reasons why authors might omit this information, but methods transparency is an important part of research reproducibility.

5 Responsible Research

Throughout the research process paper's author uphold the ethical standards of research methods and maintained the reliability of the presented results. Subsection 5.1 reviews the ethical implications of the research and Subsection 5.2 deals with the reliability of results. The author does not report any external funding or conflicts of interest.

5.1 Ethical Annotation Practices

The research assesses annotation practices in research surrounding depression. Determining the types of annotations utilized and analyzing their effectiveness in different domains provides a better understanding of the strengths and weaknesses of employed methods. Therefore the author's goal is to provide insights into ways of improvement. Allowing researchers to make informed decisions and adopt best practices in their future studies. Ultimately, by increasing the awareness and understanding of annotation practices in depression research, this research aims to foster improvements and facilitate the development of more accurate and reliable machine learning models.

5.2 Reliability of the results

During this research, certain steps were undertaken to ensure the reproducibility of obtained results. Section 2 applies PRISMA guidelines [13] and explains the used methods. The paper's author reported search terms, citation databases and applicable dates of papers collection. Additionally, data analysis and synthesis methods were detailed and the Excel sheet[6] with all the results is publicly available for further inspections.

While the author cannot ensure no information was omitted certain precautions were undertaken to minimise occurrences of negligence. To have a full overview o papers search strings were expanded when missing keywords were encountered. The author used Scopus which is one of the most robust databases available, but there is a possibility of missed papers due to the use of one citation database. An obvious limitation of the research is the presence of only one annotator examining reviewed papers. There is a possibility the author missed some information while searching for answers to my questions. To minimize overlooking important information possible answers to each question were defined to have consistent results. The results were relabeled when the understanding of the question changed.

6 Conclusions and Future Work

The research was conducted to examine and assess the quality of annotation practices in machine learning research surrounding depression. The author collected information from most cited papers, analyzed the results and provided a perspective on the quality of established practices. In section concludes the findings and provides limitations together with future work recommendations.

In this study, the author encountered different approaches to collecting annotations surrounding depression research. Annotations were sourced from human or non-human input, for both types, Section 3 details methods observed in the reviewed papers. The findings were grouped according to the research domains. The author's hypothesis was confirmed by the results: Medicine and Psychiatry papers had different methods of reporting annotation processes compared to Computer Science and Other domains.

While results for Computer Science and Other papers were found to be similar, Medicine and Psychiatry papers differed in their reporting practice. The most substantial difference was found to be in reporting of formal instructions and training. The established practice of performing research surrounding mental health is using psychometric scales [2][19], which are a form of formal instructions for the annotators.

Moreover, one way of ensuring consistency among multiple annotators is by providing formal instructions and training. Therefore Computer Science papers that tackle depression prediction would benefit from additional instruction and training for the annotators.

Even though other differences were found, some issues spanned across different domains. Across different domains, there was a lack of reporting regarding methods and data availability. This could be an issue because the transparency of the research process and data availability are both part of reproducibility guidelines. This absence of availability might have been the selected time for reviewed papers. Many reproducibility guidelines have been published in recent years [33][34], while the research reviews papers from 2013-2023. This serves as a limitation of this research.

[6] https://github.com/aleksandra-and/CSE3000-Research-Project-2022-23-Q4-

Other limitations concern mostly methods used to examine and analyze the papers. Most importantly the research was conducted by one person with little experience in psychometrics. While this could introduce bias and mistakes in recording the results, certain steps were undertaken to minimize them. The author defined the questions used to examine papers and possible answers to them. Another limitations concern study selection which consists of 80 papers sourced from one citation database. This could be mitigated by searching additional databases and expanding the analyzed set of papers.

Lastly, this research provides specific extensions that could be explored in the future. It could be interesting to examine shifts in the domain due to discussion around research reliability and transparency. This could be done by using available Excel sheet[7] and performing data analysis across years. Another possible extension of this work deals with examining the input data itself. As many people suffer from depression (4,4% [1] worldwide) it might be difficult to create comprehensive models. The author examined available information on demographics but the questions can be extended and examined in more detail.

A Included papers

Table 18: Papers that were included in the study together with the citations.

All papers included in the study					
Acharya, et al.	[35]	Kessler, et al.	[36]	Rumshisky, et al.	[37]
Acharya, et al.	[38]	Kessler, et al.	[39]	Rush, et al.	[40]
Alghowinem, et al.	[41]	Kessler, et al.	[42]	Rush, et al.	[43]
Ay, et al.	[44]	Kessler, et al.	[45]	Schalinski, et al.	[46]
Biernacka, et al.	[47]	Khodayari-Rostamabad, et al.	[48]	Scherer, et al.	[49]
Braithwaite, et al.	[50]	Koutsouleris, et al.	[51]	Schwartz, et al.	[52]
Bravo, et al.	[53]	Koutsouleris, et al.	[54]	Sharma, et al.	[55]
Burdisso, et al.	[28]	Lam, et al.	[56]	Shen, et al.	[27]
Cai, et al.	[57]	Le, et al.	[58]	Tadesse, et al.	[59]
Cai, et al.	[60]	Le, et al.	[61]	Tausczik and Pennebaker	[62]
Chekroud, et al.	[63]	Li, et al.	[64]	Tran, et al.	[65]
Chekroud, et al.	[66]	Li, et al.	[67]	Trivedi, et al.	[68]
Cheng, et al.	[69]	Liao, et al.	[70]	Trivedi, et al.	[31]
Choudhury, et al.	[30]	Lin, et al.	[71]	Tsugawa, et al.	[72]
Choudhury, et al.	[12]	Lin, et al.	[73]	Valstar, et al.	[74]
Choudhury, et al.	[75]	Losada and Crestani	[76]	Valstar, et al.	[77]
Cohn, et al.	[78]	Markowetz, et al.	[79]	van Mulligen, et al.	[80]
Cook, et al.	[81]	McCoy, et al.	[82]	Wahle, et al.	[83]
Coppersmith, et al.	[84]	Milne, et al.	[85]	Walsh, et al.	[86]
Coppersmith, et al.	[87]	Mumtaz, et al.	[88]	Walsh, et al.	[89]
Coppersmith, et al.	[90]	Mumtaz, et al.	[91]	Williamson, et al.	[92]
Deshpande and Rao	[93]	Nguyen, et al.	[94]	Wu, et al.	[95]
Dinga, et al.	[96]	O'Dea, et al.	[97]	Yates, et al.	[98]
Diniz, et al.	[99]	Orabi, et al.	[18]	Yazdavar, et al.	[100]
Donse, et al.	[101]	Passos, et al.	[102]	Zhou, et al.	[103]
Fu, et al.	[104]	Patel, et al.	[105]	Zhu, et al.	[106]
Fu, et al.	[107]	Penninx, et al.	[108]	Kessler, et al.	[29]
Hao, et al.	[32]	Perlis	[109]	Rude, et al.	[110]
Hao, et al.	[111]	Pirina and Çöltekin	[112]	Kessler, et al.	[113]
He and Cao	[114]	Preoţiuc-Pietro, et al.	[115]	Rosa, et al.	[116]
Helbich, et al.	[117]	Priya, et al.	[118]		
Hosseinifard, et al.	[119]	Ramirez-Esparza, et al.	[120]		
Iniesta, et al.	[121]	Reece, et al.	[122]		
Islam, et al.	[123]	Reece and Danforth	[124]		
Jamil, et al.	[16]	Resnik, et al.	[125]		
Jan, et al.	[126]	Resnik, et al.	[127]		
Ji, et al.	[128]	Roden, et al.	[129]		
Karstoft, et al.	[130]	Rosa, et al.	[131]		

[7] https://github.com/aleksandra-and/CSE3000-Research-Project-2022-23-Q4-

References

[1] World Health Organization. Depression and other common mental disorders: global health estimates. number-of-pages: 24.

[2] Paul E. Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Mark Simes, Richard Berman, Sarah H. Koenigsberg, and Ronald C. Kessler. The economic burden of adults with major depressive disorder in the united states (2010 and 2018). 39(6):653–665.

[3] Dominic B. Dwyer, Peter Falkai, and Nikolaos Koutsouleris. Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, 14:91–118, 5 2018.

[4] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition: DSM-IV-TR®*. Diagnostic & Statistical Manual of Mental Disorders. American Psychiatric Association.

[5] International classification of diseases (ICD).

[6] Gin S Malhi, Erica Bell, Darryl Bassett, Philip Boyce, Richard Bryant, Philip Hazell, Malcolm Hopwood, Bill Lyndon, Roger Mulder, Richard Porter, Ajeet B Singh, and Greg Murray. The 2020 royal australian and new zealand college of psychiatrists clinical practice guidelines for mood disorders. 55(1):7–117. Publisher: SAGE Publications Ltd.

[7] Stuart A. Montgomery and Marie Åsberg. A new depression scale designed to be sensitive to change. *The British Journal of Psychiatry*, 134(4):382–389, 1979.

[8] A. T. BECK, C. H. WARD, M. MENDELSON, J. MOCK, and J. ERBAUGH. An inventory for measuring depression. 4(6):561–571.

[9] Adrian B.R. Shatte, Delyse M. Hutchinson, and Samantha J. Teague. Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, 49:1426–1448, 7 2019.

[10] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? pages 325–336. Association for Computing Machinery, Inc, 1 2020.

[11] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M.B. Silenzio, and Munmun De Choudhury. A taxonomy of ethical tensions in inferring mental health states from social media. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 19:79–88, 1 2019.

[12] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 128–137. Number: 1.

[13] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. 372:n71. Publisher: British Medical Journal Publishing Group Section: Research Methods &amp; Reporting.

[14] Bibo Hao, Lin Li, Ang Li, and Tingshao Zhu. Predicting mental health status on social media a preliminary study on microblog. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8024 LNCS:101–110, 2013.

[15] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. *2nd Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2015 - Proceedings of the Workshop*, pages 31–39, 2015.

[16] Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40. Association for Computational Linguistics.

[17] Sijia Li, Yilin Wang, Jia Xue, Nan Zhao, and Tingshao Zhu. The impact of covid-19 epidemic declaration on psychological consequences: A study on active weibo users. *International Journal of Environmental Research and Public Health*, 17, 3 2020.

[18] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. Deep learning for depression detection of twitter users - ACL anthology. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97.

[19] H. J Möller. Rating depressed patients: observer- vs self-assessment. 15(3):160–172.

[20] Max Hamilton. Development of a rating scale for primary depressive illness. 6(4):278–296. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8260.1967.tb00530.x.

[21] Ulrike Kübler. Structured clinical interview for DSM-IV (SCID). In Marc D. Gellman and J. Rick Turner, editors, *Encyclopedia of Behavioral Medicine*, pages 1919–1920. Springer.

[22] Ronald C. Kessler and T. Bedirhan Üstün. The world mental health (WMH) survey initiative version of the world health organization (WHO) composite international diagnostic interview (CIDI). 13(2):93–121. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mpr.168.

[23] D. V. Sheehan, Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker, and G. C. Dunbar. The mini-international neuropsychiatric interview (m.i.n.i.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. 59 Suppl 20:22–33;quiz 34–57.

[24] Lenore Sawyer Radloff. The CES-d scale: A self-report depression scale for research in the general population. 1(3):385–401. Publisher: SAGE Publications Inc.

[25] Bernd Löwe, Jürgen Unützer, Christopher M. Callahan, Anthony J. Perkins, and Kurt Kroenke. Monitoring depression treatment outcomes with the patient health questionnaire-9. 42(12):1194–1201.

[26] Brendan Cowles and Oleg N. Medvedev. Depression, anxiety and stress scales (DASS). In Oleg N. Medvedev, Christian U. Krägeloh, Richard J. Siegert, and Nirbhay N. Singh, editors, *Handbook of Assessment in Mindfulness Research*, pages 1–15. Springer International Publishing.

[27] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression detection via harvesting social media: A multimodal dictionary learning solution. pages 3838–3844.

[28] Sergio G. Burdisso, Marcelo Errecalde, and Manuel Montes-y Gómez. A text classification framework for simple and effective early depression detection over social media streams. 133:182–197.

[29] Ronald C. Kessler, Katherine A. McGonagle, Shanyang Zhao, Christopher B. Nelson, Michael Hughes, Suzann Eshleman, Hans-Ulrich Wittchen, and Kenneth S. Kendler. Lifetime and 12-month prevalence of DSM-III-r psychiatric disorders in the united states: Results from the national comorbidity survey. 51(1):8–19.

[30] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Conference on Human Factors in Computing Systems - Proceedings*, pages 2098–2110. ISBN: 9781450333627 Publisher: Association for Computing Machinery.

[31] Madhukar H. Trivedi, Patrick J. McGrath, Maurizio Fava, Ramin V. Parsey, Benji T. Kurian, Mary L. Phillips, Maria A. Oquendo, Gerard Bruder, Diego Pizzagalli, Marisa Toups, Crystal Cooper, Phil Adams, Sarah Weyandt, David W. Morris, Bruce D. Grannemann, R. Todd Ogden, Randy Buckner, Melvin McInnis, Helena C. Kraemer, Eva Petkova, Thomas J. Carmody, and Myrna M. Weissman. Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. 78:11–23.

[32] Bibo Hao, Lin Li, Ang Li, and Tingshao Zhu. Predicting mental health status on social media. In P. L. Patrick Rau, editor, *Cross-Cultural Design. Cultural Differences in Everyday Life*, Lecture Notes in Computer Science, pages 101–110. Springer.

[33] UNESCO. UNESCO recommendation on open science. Place: 7, place de Fontenoy, 75352 Paris 07 SP, France Publisher: United Nations Educational, Scientific and Cultural Organization.

[34] Netherlands code of conduct for research integrity | NWO.

[35] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Hojjat Adeli, and D. P. Subha. Automated EEG-based screening of depression using deep convolutional neural network. 161:103–113. Publisher: Elsevier Ireland Ltd.

[36] Ronald C. Kessler, Kathleen R. Merikangas, Patricia Berglund, William W. Eaton, Doreen S. Koretz, and Ellen E. Walters. Mild disorders should not be eliminated from the DSM-v. 60(11):1117–1122.

[37] A. Rumshisky, M. Ghassemi, T. Naumann, P. Szolovits, V.M. Castro, T.H. McCoy, and R.H. Perlis. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. 6(10).

[38] U. Rajendra Acharya, Vidya K. Sudarshan, Hojjat Adeli, Jayasree Santhosh, Joel E. W. Koh, Subha D. Puthankatti, and Amir Adeli. A novel depression diagnosis index using nonlinear features in EEG signals. 74(1):79–83. Publisher: S. Karger AG.

[39] Ronald C. Kessler, Christopher H. Warner, Christopher Ivany, Maria V. Petukhova, Sherri Rose, Evelyn J. Bromet, Millard Brown, Tianxi Cai, Lisa J. Colpe, Kenneth L. Cox, Carol S. Fullerton, Stephen E. Gilman, Michael J. Gruber, Steven G. Heeringa, Lisa Lewandowski-Romps, Junlong Li, Amy M. Millikan-Bell, James A. Naifeh, Matthew K. Nock, Anthony J. Rosellini, Nancy A. Sampson, Michael Schoenbaum, Murray B. Stein, Simon Wessely, Alan M. Zaslavsky, and Robert J. Ursano. Predicting suicides after psychiatric hospitalization in US army soldiers: The army study to assess risk and resilience in servicemembers (army STARRS). 72(1):49–57. Publisher: American Medical Association.

[40] A. John Rush, Madhukar H. Trivedi, Jonathan W. Stewart, Andrew A. Nierenberg, Maurizio Fava, Benji T. Kurian, Diane Warden, David W. Morris, James F. Luther, Mustafa M. Husain, Ian A. Cook, Richard C. Shelton, Ira M. Lesser, Susan G. Kornstein, and Stephen R. Wisniewski. Combining medications to enhance depression outcomes (CO-MED): acute and long-term outcomes of a single-blind randomized study. 168(7):689–701.

[41] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parker, and Michael Breakspear. Eye movement analysis for depression detection. In *2013 IEEE International Conference on Image Processing*, pages 4220–4224. ISSN: 2381-8549.

[42] R. C. Kessler, M. B. Stein, M. V. Petukhova, P. Bliese, R. M. Bossarte, E. J. Bromet, C. S. Fullerton, S. E. Gilman, C. Ivany, L. Lewandowski-Romps, A. Millikan Bell, J. A. Naifeh, M. K. Nock, B. Y. Reis, A. J. Rosellini, N. A. Sampson, A. M. Zaslavsky, and R. J. Ursano. Predicting suicides after outpatient mental health visits in the army study to assess risk and resilience in servicemembers (army STARRS). 22(4):544–551. Number: 4 Publisher: Nature Publishing Group.

[43] A. John Rush, Maurizio Fava, Stephen R Wisniewski, Philip W Lavori, Madhukar H Trivedi, Harold A Sackeim, Michael E Thase, Andrew A Nierenberg, Frederic M Quitkin, T. Michael Kashner, David J Kupfer, Jerrold F Rosenbaum, Jonathan Alpert, Jonathan W Stewart, Patrick J McGrath, Melanie M Biggs, Kathy Shores-Wilson, Barry D Lebowitz, Louise Ritz,

George Niederehe, and for the STAR*D Investigators Group. Sequenced treatment alternatives to relieve depression (STAR*d): rationale and design. 25(1):119–142.

[44] Betul Ay, Ozal Yildirim, Muhammed Talo, Ulas Baran Baloglu, Galip Aydin, Subha D. Puthankattil, and U. Rajendra Acharya. Automated depression detection using deep representation and sequence learning with EEG signals. 43(7):205.

[45] R. C. Kessler, H. M. van Loo, K. J. Wardenaar, R. M. Bossarte, L. A. Brenner, T. Cai, D. D. Ebert, I. Hwang, J. Li, P. de Jonge, A. A. Nierenberg, M. V. Petukhova, A. J. Rosellini, N. A. Sampson, R. A. Schoevers, M. A. Wilcox, and A. M. Zaslavsky. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. 21(10):1366–1371. Number: 10 Publisher: Nature Publishing Group.

[46] Inga Schalinski, Martin H. Teicher, Daniel Nischk, Eva Hinderer, Oliver Müller, and Brigitte Rockstroh. Type and timing of adverse childhood experiences differentially affect severity of PTSD, dissociative and depressive symptoms in adult inpatients. 16(1):295.

[47] J. M. Biernacka, K. Sangkuhl, G. Jenkins, R. M. Whaley, P. Barman, A. Batzler, R. B. Altman, V. Arolt, J. Brockmöller, C. H. Chen, K. Domschke, D. K. Hall-Flavin, C. J. Hong, A. Illi, Y. Ji, O. Kampman, T. Kinoshita, E. Leinonen, Y. J. Liou, T. Mushiroda, S. Nonen, M. K. Skime, L. Wang, B. T. Baune, M. Kato, Y. L. Liu, V. Praphanphoj, J. C. Stingl, S. J. Tsai, M. Kubo, T. E. Klein, and R. Weinshilboum. The international SSRI pharmacogenomics consortium (ISPC): a genome-wide association study of antidepressant treatment response. 5(4):e553–e553. Number: 4 Publisher: Nature Publishing Group.

[48] Ahmad Khodayari-Rostamabad, James P. Reilly, Gary M. Hasey, Hubert de Bruin, and Duncan J. MacCrimmon. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. 124(10):1975–1985.

[49] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. Investigating voice quality as a speaker-independent indicator of depression and PTSD. In *Interspeech 2013*, pages 847–851. ISCA.

[50] Scott R. Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D. Barnes, and Carl Lee Hanson. Validating machine learning algorithms for twitter data against established measures of suicidality. 3(2):e4822. Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

[51] Nikolaos Koutsouleris, Christos Davatzikos, Stefan Borgwardt, Christian Gaser, Ronald Bottlender, Thomas Frodl, Peter Falkai, Anita Riecher-Rössler, Hans Jürgen Möller, Maximilian Reiser, Christos Pantelis, and Eva Meisenzahl. Accelerated brain aging in schizophrenia and beyond: A neuroanatomical marker of psychiatric disorders. 40(5):1140–1153. Publisher: Oxford Academic.

[52] H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125. Association for Computational Linguistics.

[53] Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I. Furlong. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. 16(1):55.

[54] Nikolaos Koutsouleris, Lana Kambeitz-Ilankovic, Stephan Ruhrmann, Marlene Rosen, Anne Ruef, Dominic B. Dwyer, Marco Paolini, Katharine Chisholm, Joseph Kambeitz, Theresa Haidl, André Schmidt, John Gillam, Frauke Schultze-Lutter, Peter Falkai, Maximilian Reiser, Anita Riecher-Rössler, Rachel Upthegrove, Jarmo Hietala, Raimo K. R. Salokangas, Christos Pantelis, Eva Meisenzahl, Stephen J. Wood, Dirk Beque, Paolo Brambilla, Stefan Borgwardt, and for the PRONIA Consortium. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: A multimodal, multisite machine learning analysis. 75(11):1156–1172.

[55] Manish Sharma, P. V. Achuth, Dipankar Deb, Subha D. Puthankattil, and U. Rajendra Acharya. An automated diagnosis of depression using three-channel bandwidth-duration localized wavelet filter bank with EEG signals. 52:508–520.

[56] Raymond W. Lam, Roumen Milev, Susan Rotzinger, Ana C. Andreazza, Pierre Blier, Colleen Brenner, Zafiris J. Daskalakis, Moyez Dharsee, Jonathan Downar, Kenneth R. Evans, Faranak Farzan, Jane A. Foster, Benicio N. Frey, Joseph Geraci, Peter Giacobbe, Harriet E. Feilotter, Geoffrey B. Hall, Kate L. Harkness, Stefanie Hassel, Zahinoor Ismail, Francesco Leri, Mario Liotti, Glenda M. MacQueen, Mary Pat McAndrews, Luciano Minuzzi, Daniel J. Müller, Sagar V. Parikh, Franca M. Placenza, Lena C. Quilty, Arun V. Ravindran, Tim V. Salomons, Claudio N. Soares, Stephen C. Strother, Gustavo Turecki, Anthony L. Vaccarino, Fidel Vila-Rodriguez, Sidney H. Kennedy, and on behalf of the CAN-BIND Investigator Team. Discovering biomarkers for antidepressant response: protocol from the canadian biomarker integration network in depression (CAN-BIND) and clinical characteristics of the first patient cohort. 16(1):105.

[57] Hanshu Cai, Zhidiao Qu, Zhe Li, Yi Zhang, Xiping Hu, and Bin Hu. Feature-level fusion approaches based on multimodal EEG data for depression recognition. 59:127–138.

[58] Trang T. Le, Jonathan Savitz, Hideo Suzuki, Masaya Misaki, T. Kent Teague, Bill C. White, Julie H. Marino, Graham Wiley, Patrick M. Gaffney, Wayne C. Drevets, Brett A. McKinney, and Jerzy Bodurka. Identification and replication of RNA-seq gene network modules associated with depression severity. 8(1):1–12. Publisher: Nature Publishing Group.

[59] Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. 7:44883–44893. Conference Name: IEEE Access.

[60] Hanshu Cai, Jiashuo Han, Yunfei Chen, Xiaocong Sha, Ziyang Wang, Bin Hu, Jing Yang, Lei Feng, Zhijie Ding, Yiqiang Chen, and Jürg Gutknecht. A pervasive approach to EEG-based depression detection. 2018:e5238028. Publisher: Hindawi.

[61] Trang T Le, Weixuan Fu, and Jason H Moore. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. 36(1):250–256.

[62] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. 29(1):24–54. Publisher: SAGE Publications Inc.

[63] Adam Mourad Chekroud, Ryan Joseph Zotti, Zarrar Shehzad, Ralitza Gueorguieva, Marcia K. Johnson, Madhukar H. Trivedi, Tyrone D. Cannon, John Harrison Krystal, and Philip Robert Corlett. Cross-trial prediction of treatment outcome in depression: A machine learning approach. 3(3):243–250. Publisher: Elsevier Ltd.

[64] Xiaowei Li, Bin Hu, Shuting Sun, and Hanshu Cai. EEG-based mild depressive detection using feature selection methods and classifiers. 136:151–161.

[65] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Learning vector representation of medical objects via EMR-driven nonnegative restricted boltzmann machines (eNRBM). 54:96–105.

[66] Adam M. Chekroud, Ralitza Gueorguieva, Harlan M. Krumholz, Madhukar H. Trivedi, John H. Krystal, and Gregory McCarthy. Reevaluating the efficacy and predictability of antidepressant treatments: A symptom clustering approach. 74(4):370–378.

[67] Sijia Li, Yilin Wang, Jia Xue, Nan Zhao, and Tingshao Zhu. The impact of covid-19 epidemic declaration on psychological consequences: A study on active weibo users. 17(6). Publisher: MDPI AG.

[68] Madhukar H. Trivedi, Cherise R. Chin Fatt, Manish K. Jha, Crystal M. Cooper, Joseph M. Trombello, Brittany L. Mason, Jennifer Hughes, Bharathi S. Gadad, Andrew H. Czysz, Russell T. Toll, Anne K. Fuller, Sangita Sethuram, Taryn L. Mayes, Abu Minhajuddin, Thomas Carmody, and Tracy L. Greer. Comprehensive phenotyping of depression disease trajectory and risk: Rationale and design of texas resilience against depression study (t-RAD). 122:22–32.

[69] Qijin Cheng, Tim MH Li, Chi-Leung Kwok, Tingshao Zhu, and Paul SF Yip. Assessing suicide risk and emotional distress in chinese social media: A text mining and machine learning study. 19(7):e7276. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[70] Shih-Cheng Liao, Chien-Te Wu, Hao-Chuan Huang, Wei-Teng Cheng, and Yi-Hung Liu. Major depression detection from EEG signals using kernel eigen-filter-bank common spatial patterns. 17(6):1385. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[71] Keh-Ming Lin, Hsiao-Hui Tsou, I-Ju Tsai, Mei-Chun Hsiao, Chin-Fu Hsiao, Chia-Yih Liu, Winston W Shen, Hwa-Sheng Tang, Chun-Kai Fang, Chi-Shin Wu, Shao-Chun Lu, Hsiang-Wei Kuo, Shu Chih Liu, Hsiu-Wen Chan, Ya-Ting Hsu, Jia-Ni Tian, and Yu-Li Liu. CYP1a2 genetic polymorphisms are associated with treatment response to the antidepressant paroxetine. 11(11):1535–1543. Publisher: Future Medicine.

[72] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing depression from twitter activity. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2015-April, pages 3187–3196. Association for Computing Machinery.

[73] E. Lin, P.-H. Kuo, Y.-L. Liu, Y.W.-Y. Yu, A.C. Yang, and S.-J. Tsai. A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. 9.

[74] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, AVEC '13, pages 3–10. Association for Computing Machinery.

[75] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. volume:47–56. ISBN: 9781450318891 Publisher: Association for Computing Machinery.

[76] David E. Losada and Fabio Crestani. A test collection for research on depression and language use. In Norbert Fuhr, Paulo Quaresma, Teresa Gonçalves, Birger Larsen, Krisztian Balog, Craig Macdonald, Linda Cappellato, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 28–39. Springer International Publishing.

[77] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. AVEC 2014 - 3d dimensional affect and depression recognition challenge. pages 3–10.

[78] Jeffrey F. Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De La Torre. Detecting depression from facial actions and vocal prosody. ISBN: 9781424447992.

[79] Alexander Markowetz, Konrad Błaszkiewicz, Christian Montag, Christina Switala, and Thomas E. Schlaepfer. Psycho-informatics: Big data shaping modern psychometrics. 82(4):405–411.

[80] Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. 45(5):879–884.

[81] Benjamin L. Cook, Ana M. Progovac, Pei Chen, Brian Mullin, Sherry Hou, and Enrique Baca-Garcia. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in madrid. 2016:e8708434. Publisher: Hindawi.

[82] Thomas H. McCoy, Jr, Victor M. Castro, Ashlee M. Roberson, Leslie A. Snapper, and Roy H. Perlis. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. 73(10):1064–1071.

[83] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. Mobile sensing and support for people with depression: A pilot trial in the wild. 4(3):e5960. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[84] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 shared task: Depression and PTSD on twitter. In *2nd Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2015 - Proceedings of the Workshop*, pages 31–39. ISBN: 9781941643433 Publisher: Association for Computational Linguistics (ACL).

[85] David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127. Association for Computational Linguistics.

[86] Colin G. Walsh, Jessica D. Ribeiro, and Joseph C. Franklin. Predicting risk of suicide attempts over time through machine learning. 5(3):457–469. Publisher: SAGE Publications Inc.

[87] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117. Association for Computational Linguistics.

[88] Wajid Mumtaz, Likun Xia, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, Muhammad Hussain, and Aamir Saeed Malik. Electroencephalogram (EEG)-based computer-aided technique to diagnose major depressive disorder (MDD). 31:108–115.

[89] Colin G. Walsh, Jessica D. Ribeiro, and Joseph C. Franklin. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. 59(12):1261–1270. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcpp.12916.

[90] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. pages 51–60. ISBN: 9781941643167 Publisher: Association for Computational Linguistics (ACL).

[91] Wajid Mumtaz, Likun Xia, Mohd Azhar Mohd Yasin, Syed Saad Azhar Ali, and Aamir Saeed Malik. A wavelet-based technique to predict treatment outcome for major depressive disorder. 12(2):e0171409. Publisher: Public Library of Science.

[92] James R. Williamson, Thomas F. Quatieri, Brian S. Helfer, Gregory Ciccarelli, and Daryush D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, AVEC '14, pages 65–72. Association for Computing Machinery.

[93] Mandar Deshpande and Vignesh Rao. Depression detection using emotion artificial intelligence. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 858–862.

[94] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. Affective and content analysis of online depression communities. 5(3):217–226. Conference Name: IEEE Transactions on Affective Computing.

[95] Wei Wu, Yu Zhang, Jing Jiang, Molly V. Lucas, Gregory A. Fonzo, Camarin E. Rolle, Crystal Cooper, Cherise Chin-Fatt, Noralie Krepel, Carena A. Cornelssen, Rachael Wright, Russell T. Toll, Hersh M. Trivedi, Karen Monuszko, Trevor L. Caudle, Kamron Sarhadi, Manish K. Jha, Joseph M. Trombello, Thilo Deckersbach, Phil Adams, Patrick J. McGrath, Myrna M. Weissman, Maurizio Fava, Diego A. Pizzagalli, Martijn Arns, Madhukar H. Trivedi, and

Amit Etkin. An electroencephalographic signature predicts antidepressant response in major depression. 38(4):439–447. Number: 4 Publisher: Nature Publishing Group.

[96] Richard Dinga, Lianne Schmaal, Brenda W. J. H. Penninx, Marie Jose van Tol, Dick J. Veltman, Laura van Velzen, Maarten Mennes, Nic J. A. van der Wee, and Andre F. Marquand. Evaluating the evidence for biotypes of depression: Methodological replication and extension of drysdale et al. (2017). 22:101796.

[97] Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. Detecting suicidality on twitter. 2(2):183–188. Publisher: Elsevier.

[98] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978. Association for Computational Linguistics.

[99] B. S. Diniz, E. Sibille, Y. Ding, G. Tseng, H. J. Aizenstein, F. Lotrich, J. T. Becker, O. L. Lopez, M. T. Lotze, W. E. Klunk, C. F. Reynolds, and M. A. Butters. Plasma biosignature and brain pathology related to persistent cognitive impairment in late-life depression. 20(5):594–601. Number: 5 Publisher: Nature Publishing Group.

[100] Amir Hossein Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 1191–1198. Association for Computing Machinery.

[101] Lana Donse, Frank Padberg, Alexander T. Sack, A. John Rush, and Martijn Arns. Simultaneous rTMS and psychotherapy in major depressive disorder: Clinical outcomes and predictors from a large naturalistic study. 11(2):337–345.

[102] Ives Cavalcante Passos, Benson Mwangi, Bo Cao, Jane E. Hamilton, Mon-Ju Wu, Xiang Yang Zhang, Giovana B. Zunta-Soares, Joao Quevedo, Marcia Kauer-Sant'Anna, Flávio Kapczinski, and Jair C. Soares. Identifying a clinical signature of suicidality among patients with mood disorders: A pilot study using a machine learning approach. 193:109–116.

[103] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20k dataset.

[104] Cynthia H. Y. Fu, Steven C. R. Williams, Anthony J. Cleare, Michael J. Brammer, Nicholas D. Walsh, Jieun Kim, Chris M. Andrew, Emilio Merlo Pich, Pauline M. Williams, Laurence J. Reed, Martina T. Mitterschiffthaler, John Suckling, and Edward T. Bullmore. Attenuation of the neural response to sad faces in major depressionby antidepressant treatment: A prospective, event-related functional magnetic resonance ImagingStudy. 61(9):877–889.

[105] Meenal J. Patel, Carmen Andreescu, Julie C. Price, Kathryn L. Edelman, Charles F. Reynolds III, and Howard J. Aizenstein. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. 30(10):1056–1067. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gps.4262.

[106] Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics. 9(4):578–584. Conference Name: IEEE Transactions on Affective Computing.

[107] Cynthia H. Y. Fu, Janaina Mourao-Miranda, Sergi G. Costafreda, Akash Khanna, Andre F. Marquand, Steve C. R. Williams, and Michael J. Brammer. Pattern classification of sad facial processing: Toward the development of neurobiological markers in depression. 63(7):656–662.

[108] Brenda W.J.H. Penninx, Aartjan T.F. Beekman, Johannes H. Smit, Frans G. Zitman, Willem A. Nolen, Philip Spinhoven, Pim Cuijpers, Peter J. De Jong, Harm W.J. Van Marwijk, Willem J.J. Assendelft, Klaas Van Der Meer, Peter Verhaak, Michel Wensing, Ron De Graaf, Witte J. Hoogendijk, Johan Ormel, and Richard Van Dyck. The netherlands study of depression and anxiety (NESDA): rationale, objectives and methods. 17(3):121–140. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mpr.256.

[109] Roy H. Perlis. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. 74(1):7–14.

[110] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. Language use of depressed and depression-vulnerable college students. 18(8):1121–1133. Publisher: Routledge _eprint: https://doi.org/10.1080/02699930441000030.

[111] Bibo Hao, Lin Li, Ang Li, and Tingshao Zhu. Predicting mental health status on social media a preliminary study on microblog. 8024 LNCS:101–110. ISBN: 9783642391361 Publisher: Springer Verlag.

[112] Inna Pirina and Çağrı Çöltekin. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12. Association for Computational Linguistics.

[113] Ronald C. Kessler, Irving Hwang, Claire A. Hoffmire, John F. McCarthy, Maria V. Petukhova, Anthony J. Rosellini, Nancy A. Sampson, Alexandra L. Schneider, Paul A. Bradley, Ira R. Katz, Caitlin Thompson, and Robert M. Bossarte. Developing a practical suicide risk prediction model for targeting high-risk patients in the veterans health administration. 26(3):e1575. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mpr.1575.

[114] Lang He and Cui Cao. Automated depression analysis using convolutional neural networks from speech. 83:103–111.

[115] Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H. Andrew Schwartz, and Lyle Ungar. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30. Association for Computational Linguistics.

[116] Maria J. Rosa, Liana Portugal, Tim Hahn, Andreas J. Fallgatter, Marta I. Garrido, John Shawe-Taylor, and Janaina Mourao-Miranda. Sparse network-based models for patient classification using fMRI. 105:493–506.

[117] Marco Helbich, Yao Yao, Ye Liu, Jinbao Zhang, Penghua Liu, and Ruoyu Wang. Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in beijing, china. 126:107–117. Publisher: Pergamon.

[118] Anu Priya, Shruti Garg, and Neha Prerna Tigga. Predicting anxiety, depression and stress in modern life using machine learning algorithms. 167:1258–1267.

[119] Behshad Hosseinifard, Mohammad Hassan Moradi, and Reza Rostami. Classifying depression patients and normal subjects using machine learning techniques and nonlinear features from EEG signal. 109(3):339–345. Publisher: Elsevier.

[120] Nairan Ramirez-Esparza, Cindy Chung, Ewa Kacewic, and James Pennebaker. The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches. 2(1):102–108. Number: 1.

[121] Raquel Iniesta, Karim Malki, Wolfgang Maier, Marcella Rietschel, Ole Mors, Joanna Hauser, Neven Henigsberg, Mojca Zvezdana Dernovsek, Daniel Souery, Daniel Stahl, Richard Dobson, Katherine J. Aitchison, Anne Farmer, Cathryn M. Lewis, Peter McGuffin, and Rudolf Uher. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. 78:94–102.

[122] Andrew G. Reece, Andrew J. Reagan, Katharina L. M. Lix, Peter Sheridan Dodds, Christopher M. Danforth, and Ellen J. Langer. Forecasting the onset and course of mental illness with twitter data. 7(1):1–11. Publisher: Nature Publishing Group.

[123] Md. Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang, and Anwaar Ulhaq. Depression detection from social network data using machine learning techniques. 6(1):8.

[124] Andrew G. Reece and Christopher M. Danforth. Instagram photos reveal predictive markers of depression. 6(1). Publisher: SpringerOpen.

[125] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond LDA: Exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107. Association for Computational Linguistics.

[126] Asim Jan, Hongying Meng, Yona Falinie Binti A. Gaus, and Fan Zhang. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. 10(3):668–680. Conference Name: IEEE Transactions on Cognitive and Developmental Systems.

[127] Philip Resnik, Anderson Garron, and Rebecca Resnik. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353. Association for Computational Linguistics.

[128] Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. Supervised learning for suicidal ideation detection in online user content. 2018:e6157249. Publisher: Hindawi.

[129] D. M. Roden, J. M. Pulley, M. A. Basford, G. R. Bernard, E. W. Clayton, J. R. Balser, and D. R. Masys. Development of a large-scale de-identified DNA biobank to enable personalized medicine. 84(3):362–369.

[130] Karen Inge Karstoft, Isaac R. Galatzer-Levy, Alexander Statnikov, Zhiguo Li, Arieh Y. Shalev, Yael Ankri, Sara Freedman, Rhonda Addesky, Yossi Israeli-Shalev, Moran Gilad, and Pablo Roitman. Bridging a translational gap: Using machine learning to improve the prediction of PTSD. 15(1). Publisher: BioMed Central Ltd.

[131] Renata Lopes Rosa, Gisele Maria Schwartz, Wilson Vicente Ruggiero, and Demóstenes Zegarra Rodríguez. A knowledge-based recommendation system that includes sentiment analysis and deep learning. 15(4):2124–2135. Conference Name: IEEE Transactions on Industrial Informatics.