PLOS ONE

RESEARCH ARTICLE

Assessing ML classification algorithms and NLP techniques for depression detection: An experimental case study

Giuliano Lorenzoni*, Cristina Tavares, Nathalia Nascimento, Paulo Alencar, Donald Cowan

David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada

* glorenzo@uwaterloo.ca

Check for updates

OPEN ACCESS

Citation: Lorenzoni G, Tavares C, Nascimento N, Alencar P, Cowan D (2025) Assessing ML classification algorithms and NLP techniques for depression detection: An experimental case study. PLoS One 20(5): e0322299. https://doi.org/10.1371/journal.pone.0322299

Editor: Asim Mehmood, Jazan University, SAUDI ARABIA

Received: April 26, 2024

Accepted: March 18, 2025

Published: May 28, 2025

Copyright: © 2025 Lorenzoni et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: DAIC-WOZ database is publicly available to the research community only upon request (https://dcapswoz.ict.usc.edu/).

Funding: This study was supported by grants awarded to PA and DC by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Centre for Community Mapping (COMAP).

Competing interests: The authors have declared that no competing interests exist.

Abstract

Context and background. Depression has affected millions of people worldwide and has become one of the most common mental disorders. Early mental disorder detection can reduce costs for public health agencies and prevent other major comorbidities. Additionally, the shortage of specialized personnel is very concerning since depression diagnosis is highly dependent on expert professionals and is time-consuming. **Research problems**. Recent research has evidenced that machine learning (ML) and natural language processing (NLP) tools and techniques have significantly benefited the diagnosis of depression. However, there are still several challenges in the assessment of depression detection approaches in which other conditions such as post-traumatic stress disorder (PTSD) are present. These challenges include assessing alternatives in terms of data cleaning and pre-processing techniques, feature selection, and appropriate ML classification algorithms. **Purpose of the study**. This paper tackles such an assessment based on a case study that compares different ML classifiers, specifically in terms of data cleaning and pre-processing, feature selection, parameter setting, and model choices. **Methodology**. The experimental case study is based on the Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ) dataset, which is designed to support the diagnosis of mental disorders such as depression, anxiety, and PTSD. **Major findings**. Besides the assessment of alternative techniques, we were able to build models with accuracy levels around 84% with Random Forest and XGBoost models, which is significantly higher than the results from the comparable literature which presented the level of accuracy of 72% from the SVM model. **Conclusions**. More comprehensive assessments of ML classification algorithms and NLP techniques for depression detection can advance the state of the art in terms of improved experimental settings and performance.

1 Introduction

Depression has affected millions of people worldwide. It is considered a common mental disorder not only because of the increasing number of cases but also because of its intrinsic relation to other psychiatric and physical illnesses such as stroke, heart disease, and chronic

fatigue. The effects of the pandemic on general mental health, the recent rise in cases of mental health issues, and the shortage of professionals specialized in the diagnosis and treatment of mental disorders such as depression all characterize a serious issue that can have several negative implications for society.

In this context, one way to help to cope with the effects of the rising number of depression cases and the shortage of medical specialists is to introduce early detection methods. Recent research has evidenced that machine learning (ML) and natural language processing (NLP) tools and techniques can significantly benefit the early diagnosis of depression. However, there are still several challenges in the assessment of depression detection approaches in which other conditions such as post-traumatic stress disorder (PTSD) are present. These challenges include assessing alternatives in terms of data cleaning and pre-processing techniques, feature selection, and appropriate ML classification algorithms.

This paper tackles such an assessment based on a case study that compares different ML classifiers, specifically in terms of data cleaning and pre-processing, feature selection, parameter setting, and model selection. The case study is based on the Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ) dataset [1], which is designed to support the diagnosis of mental disorders such as depression, anxiety, and PTSD. Besides the assessment of alternative techniques, we were able to build models with accuracy levels around 84% with Random Forest and XGBoost models, which is significantly higher than the results from the comparable literature which presented the level of accuracy of 72% from the SVM model.

The proposed approach advances the state of the art in terms of improved experimental settings and performance. In this sense, we follow a comprehensive experiment workflow to select the optimum model. Ths workflow allows to explore distinct preprocessing technologies, multiple feature combinations, varied parameters, and different selected estimators. Exploring different aspects based on an explicit workflow helped us to reach perfomance improvements in the case study.

This paper is structured as follows. Sect 2 presents related work. Sect 3 describes our experiment design and Sect 4 presents the results of our study. Sect 5 provides some discussion on insights about the models, selected features, and other factors and their dependencies. Sect 6 describes threats to validity, and, finally, Sect 7 presents our conclusions and future work.

2 Related work

Regarding depression detection in general, recent studies have shown that machine learning tools can be applied to predict mental health diseases. Saqib et al. conducted a literature review to investigate the use of machine learning in the prediction of postpartum depression (PPD) and found 14 research works in this theme. Findings revealed progressive results in performance of machine learning methods and techniques that could benefit early PPD detection significantly [2]. A meta-analysis and systematic review conducted by Lee et al. examined the use of machine learning algorithms to predict mood disorders and support on the selection for treatment and therapy. Authors provided an overview of machine learning applications applied to population with mood disorders. The study indicated high accuracy results in the the prediction of therapeutic outcomes, and conclusions suggested that machine learning techniques could be powerful tools in supporting the research of mood disorders and analysis of multiple types of data from diverse sources [3].

Spacic and Goran have conducted a study to investigate the use and application of text data on clinical practice using machine learning and NLP techniques [4] The study was based on a systematic literature review aiming to collect information about research related these knowledge areas. Findings revealed major barriers to use machine learning for clinical practice such

as a lack of use of large datasets to train models and difficulties in the annotation process. The authors' conclusions suggested further research in unsupervised learning or data augmentation and transfer learning. Finally, although we have identified related literature reviews and practical applications, there is still a need to advance the application of techniques to support depression detection models and improve their accuracy.

In terms of the ways NLP techniques can be applied in the study of clinical depression, for example, using social media data or other forms of text data, the related efforts falls into the following categories:

- The articles which compare different techniques in order to find the best model according to some evaluation metrics, such as [5–13].
- The articles that proposed new detection or prediction models, platforms, or systems, namely [13–21] or articles regarding specific applications [15,22–30].
- The articles focused on advanced feature selection techniques based on nature-inspired computing algorithms, applied to disease prediction [31–34].

Specifically, regarding the articles from the third group, some authors [34] introduce an innovative approach leveraging the Ant-Lion Optimization algorithm to select relevant features from large datasets, applied to chronic diseases such as diabetes, diabetic retinopathy, skin cancer, and heart disease. In [32] the researchers optimize feature selection in fundus images using a hybrid of the Emperor Penguin Optimization (EPO) and Bacterial Foraging Optimization (BFO) algorithms, achieving high accuracy in glaucoma detection. The authors of [31] propose a hybrid approach combining the Gravitational Search Algorithm (GSA) and EPO for effective binary breast cancer classification. Finally, in [33] a similar strategy using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset focuses on improving model generalization and providing reliable clinical diagnostic support. Although feature selection is a pivotal part of our work, these articles are not directly comparable because they are derived from numerical data, such as those from bioinformatics, medical images, and features extracted from biomedical datasets. However, these techniques may be adapted and applied to text data or NLP features calculated from a dataset like DAIC-WOZ, which consists of interviews with subjects.

We have also found several articles which belong to more than one group. For instance, we have found articles proposing a new model and testing it along with other techniques, such as [35] whose authors proposed a system to detect depression using NLP techniques and compared the performance of this method with multiple machine learning algorithms in order to identify best-performing configuration.

Similarly, article [25] presents a machine learning framework for the automatic detection and classification of 15 common cognitive distortions (defined by the authors as automatic and self-reinforcing irrational thought patterns) in two novel mental health free datasets collected from both crowdsourcing and a real-world online therapy program, comparing the results of such framework with multiple traditional machine learning models/techniques (e.g XG-Boost, SVM, CNN, and RNN). Likewise, the authors of [36] present an automated conversational platform that was used as a preliminary method of identifying depression associated risks and have its results compared with the ones provided by the classic SVM classifier. Moreover, this article goes further on the classification task and looks for the identification of different levels of states ('happy', 'neutral', 'depressive' and 'suicidal') like the work of the authors of [24] which classifies different degrees (or categories) of depression (Minimal, Mild, Moderate, Severe). There are also articles on specific applications involving the proposal of new detection models, systems, or platforms in a specific context. For example, in [15], the

authors present DEPA, which is a self-supervised, pre-trained depression audio embedding method for depression detection, and also explore self-supervised learning in a specific task within audio processing.

Regarding ML models and techniques applied in depression studies, two main groups of studies emerge:

- First group: Articles whose techniques were applied in studies comparing traditional machine learning models to identify the best model according to some evaluation metrics (i.e. accuracy), such as in [5–12].
- Second group: Articles proposing new techniques for depression detection in a specific context [13–18,23–26,37]. These techniques were compared with the traditional models.

As for the techniques mentioned in the articles from the first group, we can highlight: (a) Support Vector Machine (SVM) [5–12]: (b) Random Forest (RF) [5–8,10–12]: (c) Gradient Boosted Trees (XGBoost) [6,12]: (d) Naive Byes Classifiier [6,7,11]: (e) Logistic Regression (LR) [5,6,11,12]: (f) Decision Tree (DT) [5,6,8,11,12]: (g) Extra Trees Classifier [7]: h) Stochastic Gradient Boost Classifier [7]: (i) Long short-term memory (LSTM)/ recurrent neural network (RNN) [8]: (j) Ensemble methods [8]: (k) Multilayer Perceptron Classifier (MLP) [9]: (l) Multinomial Naive Bayes (NB) [11]: and (m) K-Nearest Neighbor (KNN) [11].

Along with most of the traditional machine learning models mentioned in the articles from the first group, it is worth mentioning the following techniques used in the articles in the second group: a) Deep neural network classification model - Multimodal Feature Fusion Network (MFFN) [13]: b) Bidirectional Encoder Representations of Transformers (BERT); and c) Robustly optimized BERT approach (RoBERTa). Still, regarding the second group, we have identified the use of some transformer models [37], which are semi-supervised machine learning techniques that are primarily used with text data, and constitute an alternative to recurrent neural networks in natural language processing tasks. Basically, it consists in deep learning models that adopt the mechanism of self-attention, deferentially weighting the significance of each part of the input data. It is used primarily in the field of natural language processing (NLP) and in computer vision (CV).

In summary, although we also found some unsupervised models among the articles in the second group, most of the work in the literature seems to focus on supervised approaches such as Support Vector Machines, Random Forest, and XGBoost. Finally, Despite these advancements, further research is required to refine depression detection models and enhance their precision.

In another example of a representative article from the second group, in [38] the authors evaluated four different classifiers (XGBoost, Random Forest, Logistic Regression, and SVM) using textual data extracted from social media (i.e., Twitter) to predict depression levels, performing direct comparisons of accuracy and computational efficiency between the models. Similarly to our study, the authors conducted data cleaning and feature extraction steps, including the removal of stop words, stemming, and lemmatization, to preprocess the textual data for the models. However, beyond the difference in data sources, differences were also observed in the types of data analyzed. While our study focuses on textual data derived from transcribed interviews, exploring features such as average response time, word frequency, and sentiment metrics specific to each question, these authors [38] analyzed general sentiments and keywords in social media posts, placing less emphasis on introducing more complex features.

Specifically regarding the articles using the DAIC-WOZ database the detection of depression through ML and NLP has presented varied approaches, with a significant body of literature exploring this domain. The methodologies range from text analysis on social media [39] to speech emotion analysis [40], highlighting the multidisciplinary nature of the field. In [41], the authors trained ML models on textual data from AVEC-19, a subset of the 160 DAIC-WOZ corpus, to identify individuals with PTSD through sentiment analysis of semi-structured interviews, achieving a balanced accuracy of 80.4%. Furthermore, they implemented various partitioning techniques to evaluate the generalizability of the model, demonstrating that it is possible to use sentiment analysis in speech to identify the presence of PTSD, even in a virtual environment. In contrast to the approach described in [41] our study addresses a variety of algorithm combinations and provides an exhaustive investigation of feature selection and hyperparameter tuning. Additionally, while the authors in [41] address sentiment analysis features, our work expands the analysis by incorporating aspects such as word frequency and speech-based metrics.

Text-based studies, like those conducted by Liu et al. [39] and Chaves et al. [42], focus on detecting depressive symptoms using sentiment analysis and summarization techniques on social media platforms. These studies emphasize the potential of text mining in identifying depression but also recognize the challenges of data bias and privacy concerns. In contrast, speech-based analyses, as can be found in the works of Bhavya et al. [40], Squires et al. [43], and Cummins et al. [44], investigate acoustic features and speech patterns using datasets like DAIC-WOZ for emotion recognition, presenting the challenge of data scarcity and the need for robust and generalizable models.

Studies like Krishna and Anju [45] and Nguyen et al. [46] extend the exploration into multimodal domains, arguing for the combination of audiovisual cues with linguistic analysis. They highlight the potential of deep learning models to integrate these diverse data streams for higher diagnostic accuracy. This trend is echoed by Nouman et al. [47], which explores contactless sensing technologies, underscoring the importance of multifaceted data in enhancing predictive models.

The gaps in the literature are prominently presented, with a recurring theme being the scarcity of large and diversified data sets, a challenge that Silva et al. [48] and Yan et al. [49] address by calling for open-access resources to bolster research efforts, tackling everything from data scarcity to ethical dilemmas. More specifically, these studies offer comprehensive reviews, the former focusing on computational methods and databases and the latter on the broader challenges that AI faces in recognizing mental disorders (for example, addressing logical fallacies and diagnostic criteria that often hinder AI's potential in mental health). These reviews also highlight the gaps in standardization and the need for longitudinal analyses to advance the field.

Similarly, Song et al. [50] and Janardhan and Nandshini [51] emphasize the role of deep learning and feature selection in improving the accuracy of depression prediction, pointing out the limitations of small data sets and advocating for larger and more diverse data collections.

Additionally some of the mentioned studies (such as [48] and [49]) methodologically reveal a preference for ensemble and hybrid ML models. This information is also evidenced in the works of Mao et al. [52] and Janardhan and Nandhini [51], which prioritize the combination of classifiers for improved predictive outcomes. In fact, Mao et al. [52] provides a systematic review of automated clinical depression diagnosis, synthesizing findings across 264 studies. The review emphasizes the importance of acoustic features and their correlation with observable symptoms, while also pointing out methodological variations and ethical considerations that need to be addressed in future research.

The research also points to a growing dependence on sophisticated data preprocessing and feature extraction techniques, as detailed in Aleem et al. [53] and Nouman et al. [47], to mitigate the problems of class imbalance and data scarcity. Indeed, Aleem et al. [53] and Ndaba et al. [54] discuss the complexities of data handling techniques for depression prediction, addressing the problem of class imbalance prevalent in health records. They suggest that future work should explore under-sampling techniques and regression modeling for a more nuanced approach to depression detection.

Finally, ethical concerns, particularly in studies leveraging social media data for depression detection, are also a significant consideration, necessitating a delicate balance between innovation and privacy, as highlighted in [39] and [49].

In summary, the kaleidoscope of methodologies ranges from the traditional - decision trees and SVMs, explored in Yang, Jiang, and He [55] and Yang et al. [56], to the avant-garde - deep learning architectures that carve out nuanced patterns within multimodal data sets, as highlighted in Squires et al. [43] and Krishna and Anju [45]. These methodologies are not without their challenges, as elucidated by Cummins et al. [44], where the limitations of dataset sizes and deep learning's voracious appetite for data become apparent.

Although these articles use the same dataset used in our work, we were able to spot several differences. Our work, utilizes the DAIC-WOZ database predominantly for PTSD patients, implementing sentiment analysis within NLP to analyze clinical interviews. In contrast, the approach of Liu et al. [39] capitalizes on the expanse of social media, engaging with a more volatile and expansive data collection environment. Our methodology provides a narrower scope compared to Liu et al.'s review of diverse social platforms for depressive symptom detection, emphasizing sentiment analysis tailored to PTSD. In the realm of speech emotion analysis, while Bhavya et al. [40] narrows down to depression recognition via speech emotion analysis using the same database, our work extends the analysis to encompass a broader spectrum of mental disorders. This comprehensive approach allows us to evaluate classifier performance across different mental health conditions, including depression, anxiety, and PTSD. Additionally, our work remains grounded in the domain of NLP tasks related to depression, utilizing ML classifiers, whereas Krishna and Anju [45] and Nguyen et al. [46] advocate for a multimodal approach. This highlights our specific focus on linguistic analysis as a powerful tool for mental health diagnostics, despite recognizing the value of multimodal approaches.

When addressing the challenges of dataset availability and ethical considerations, our work benefits from an established dataset, allowing us to sidestep broader ethical debates and focus on the optimization of classifiers. On the other hand, Silva et al. [48] and Yan et al. [49] delve into these challenges, highlighting the ethical complexities inherent in AI for mental health, a perspective that enriches the dialogue within the field.

Furthermore, Aleem et al. [53] and Ndaba et al. [54] discuss class imbalance and feature engineering, proposing solutions like under-sampling techniques and regression modeling. While their contributions to the field are significant, our research primarily centers on the efficacy of various ML classifiers, showcasing the strengths of Random Forest and XGBoost.

Mao et al. [52] provides a systematic review of automated clinical depression diagnosis, encompassing a broad range of studies and datasets. In contrast, our work offers a focused comparison of classifiers on a singular dataset, providing a detailed lens on classifier performance in PTSD-related depression detection.

Lastly, our work does not cover the advances presented by Nouman2022's exploration of contactless sensing technologies [47], nor does it delve into the deep learning for psychiatric applications as investigated by Squires et al. [43]. These studies present expanded views and forward-thinking perspectives on mental health monitoring and data analysis tools that differ from the direct application of specific classifiers evaluated in our work.

In this way, the most closely related works to our paper can be found among the articles utilizing advanced machine learning techniques and the DAIC-WOZ database for mental health diagnostics. Still, the work of Yang et al. [56] employs methodologies akin to ours, such as Support Vector Machines (SVMs) and Text Convolutional Neural Network (TextCNN) for text-based emotional analysis, reflecting a similar application of sophisticated ML classifiers within the NLP framework. This parallel approach offers a complementary viewpoint to our work, enhancing the understanding of depression detection through linguistic cues and semantic analysis. The insights from [56], particularly the application of SVM and TextCNN, are invaluable for refining current practices and bolstering the predictive power of ML models in the domain of mental health, drawing a more comprehensive map of the current landscape and future directions in depression detection using machine learning and natural language processing.

3 Experiment design

The objective of this case study is to build a diagnostic model for depression disorders based on different supervised ML models and NLP techniques. We explored multiple model tuning configurations, feature sets, and data preprocessing methodologies across all the models incorporated in our analysis.

As a result, the design of this case study was conceived to identify the optimal model, the most suitable parameters, and any combination of factors that could provide the best results in terms of accuracy subject to achieving the highest possible F1-Score given the data imbalance. We also established a baseline approach to check the efficiency of the models and examine the insights of our case study based on our results.

Fig 1 showcases a comprehensive diagrammatic representation of the overall workflow we adopted to develop this experiment. This workflow is inspired by the one proposed by Amershi et al. [57]. Within this workflow, we delineated the steps that influenced the model selection. For instance, during the training phase, we explored two distinct preprocessing techniques, multiple feature combinations, varied parameters, and different selected estimators [58] until we surpassed the baseline's performance. Subsequently, during the model evaluation phase, we assessed the final selected model using the safeguard dataset, which was previously separated during the data splitting phase.

As shown in the Fig 1, the final model selection is based on the chosen algorithm, a specific parameter configuration, and a set of features. Any changes in the algorithm, parameter tuning, or feature sets are regarded as distinct models (classifiers), each with its own results.

The next subsections describe the implementation details of each stage, tailored for this specific case study.

3.1 Data collection: Dataset

We based our research on the Distress Analysis Interview Corpus - Wizard-of-Oz (DAIC-WOZ), a dataset designed to support the diagnosis of mental disorders such as depression, anxiety, and post-traumatic stress disorder. DAIC is a database that is part of a larger corpus [59]. The DAIC-WOZ database is publicly available to the research community upon request at https://dcapswoz.ict.usc.edu/.

We have submitted a request and the data released refers only to the depressed patients' database. However, due to ethical and legal restrictions associated with the dataset, we do not have permission to share the data directly. Researchers interested in replicating our study can obtain access to the DAIC-WOZ dataset by following the official request process outlined



Fig 1. Diagrammatic representation of the overall workflow of our experiment to select the optimum model. The red arrow illustrates that model training may loop back to estimator selection [58], pre-processing, feature engineering, and parameter setting.

https://doi.org/10.1371/journal.pone.0322299.g001

by its custodians. This ensures compliance with data privacy regulations while maintaining scientific transparency.

This database contains clinical interviews conducted by humans, human-controlled agents, and autonomous agents. The computer agent is an animated virtual interviewer robot called Ellie that identifies mental illness indicators. Data includes 189 sessions of interviews which correspond to questionnaire responses and audio and video recordings. Each interview is identified by a session number and has a correspondent folder of files which includes files related to video features, audio, transcript, and audio features. This dataset contains training, development, and test subsets. Interviewees are both distressed and non-distressed individuals.

Specifically, the PHQ8_Score file includes the following fields:

- The identification of the participant associated with the text record (Participant_ID);
- A binary item (PHQ8_Binary) that indicates whether the participant has depression or not (0 is depression, and 1 is no depression);
- A score that indicates how severe is the depression (PHQ8_Score) (PHQ8_Score > 10 corresponds to PHQ8_Binary=1); and
- The gender of the participant, which was not used in the experiment (Gender).

As an illustration, Table 2 shows some of the data rows (from [59]).

We used the text files related to the transcript of the interviews. The Patient Health Questionnaire depression scale (PHQ-8) defines the depression diagnostic and severity measure. and the PHQ-8 file which described the score of each patient according to the depression scale described in [60].

3.2 Baseline investigation

Before determining what techniques will be used in our case study we established a baseline approach. Because our classifier will choose among depressed and not depressed states (0 or 1), our approach consists in setting the same state (e.g. depressed, that is 0) to each observation of our training set and then checking the error rate in the predictions of our testing set. The purpose is to compare the accuracy rate of our chosen models not only amongst themselves but also with the baseline approach to gauge the model's efficacy.

3.3 Data splitting: Training and testing set

For this study, we will use approximately 80% of our data set as a training set (150 interview transcripts) and the remaining approximately 20% (38 interview transcripts) as our testing set.

Because three meetings were removed from the data set due to structural problems, we ended up with 148 interviews in our training set and 37 in our testing set.

3.4 Estimator selection

As outlined in [61], certain heuristics enable the pre-selection of machine learning estimators [58] that are best suited for a particular dataset or problem. By doing so, during the pursuit of the optimal model, it is possible to narrow down a group of most compatible estimators, thereby reducing our search options. Given the characteristics of the chosen problem and dataset (including its labeled nature, number of labels, size, number of features, and application output), and the algorithms that we identified in the related work section, our exploration was confined to three machine-learning techniques. They are:

- Random Forest Classifier (RF): an algorithm that builds multiple decision trees and makes a final prediction based on the majority outcome of each decision tree.
- XGBoost (XGB): an algorithm that sequentially builds decision trees that recurrently diminish the error of the previous prediction by applying gradient descent on a loss function when adding new trees.
- Support Vector Machine (SVM): an algorithm that maps the data points to another space (using kernel algorithms) so that it is possible to separate (and therefore classify) the data points into different categories.

3.5 Pre-processing: Data cleaning

Different attempts were made to clean the data. Initially, a custom set of stopwords was removed from the conversation, but the end result of the classification came out worse in this scenario as opposed to when no words were extracted. Moreover, the conversations had basic punctuation removed and the letters converted to lowercase. Afterward, the artificial conversation markers in the dataset (which did not represent words actually spoken in the dialogue, but rather indicated what kind of comment was being made) were removed to keep the text as close as possible to real conversations. Finally, the datasets that only contained interviewee speeches (and no comments or questions from the bot) were not used, given that most of the features in the model relied on questions asked by the bot.

3.6 Feature engineering: Features selection

Although the objective is to test a different set of features in our chosen NLP algorithm, it is important to mention that some of our features, such as the Sentiment Analysis calculation and the average of unique words or stop words, were also derived from other NLP techniques.

Table 1. Explanation of features.

Feature	Description
Answers sentiments as Features	The dataset contains a transcription of a conversation between a bot and a person, during which the bot asks its counterpart several questions. The sentiment of the answers to the most relevant questions are used as features in our model.
Sentiment of an Interview (avg sentiment)	Sentiment of all the comments made by the person throughout the conversation with the bot.
Average Response Time (avg response time)	Average time for the interviewee to answer a question from the bot.
Speech speed	Calculated by measuring the number of words for each comment of the interviewee, measuring the time it took to make that comment, dividing those two values, and taking the average of that result for all comments made by the interviewee throughout the conversation.
Average Unique Words Frequency (Avg unique frequency)	Calculated by measuring the number of unique (without counting repetition) words spoken in a comment, then measuring the number of words in that comment, dividing the first by the second, and taking the average of that calculation for all comments made by the interviewee throughout the conversation.
Average Stop Words Frequency (Avg sw frequency)	Calculated similarly to the previous metric (avg unique frequency), but replacing unique words with stop words, which are words without expressive content (e.g., "uh", "um", "mm").
Average Number of Characters (Avg characters)	Calculated by measuring the number of characters of each comment, dividing that value by the number of words in that comment, and then averaging that calculation through all comments made by the interviewee in the meeting.
Average Noun, Verbs, Adjectives, Adverbs (Avg nouns, avg verbs, adj freq, avg adv)	Calculated similarly to the stop word frequency (Avg sw frequency), but instead of stop words, it calculates the average frequency of nouns, verbs, adjectives and adverbs, respectively.
Average of First-Person Related Words (fp avg)	Similar to stop word frequency (Avg sw frequency), but instead of stop words, it measures the frequency of first-person-related words (e.g. "I", "we", "us").

https://doi.org/10.1371/journal.pone.0322299.t001

Table 2. PHQ8_Score File.

Participant_ID	PHQ8_Binary	PHQ8_Score	Gender
302	0	4	1
335	1	12	0
367	1	19	1
377	1	16	0

https://doi.org/10.1371/journal.pone.0322299.t002

We have built a total of 30 features (since the first feature described contains 19 separated features) which were combined in different ways and applied across the 3 NLP classifiers to find the best combination in terms of model accuracy. The list of the features considered on this study and its explanation can be found in Table 1.

Finally, it is also worth mentioning that the best results from our case study were obtained after testing several combinations of features from this list (and even features that were originally considered but then discarded because of poor results) including the sentiment of all the interviews and the sentiment of the answers from some selected questions that appeared in most of the transcripts. Additional information about the features selection can be found in the Results section.

3.7 Parameters setting

Random forest classifier. This classifier has 18 parameters, out of which max depth and random state were set to non-default values. Random state is used to promote the reproducibility of the model, whereas max depth influences overfitting tendencies, which is why it was included in the training process.

XGBoost. This classifier has 33 parameters, out of which n estimators, eval metric, learning rate, label encoder, max depth, and random state were set to non-default values. Random state is used to promote the reproducibility of the model, whereas the max depth influences the overfitting tendency and is thus included in the training process. 27 of the parameters for this model are optional.

Support vector machine. This classifier has 15 parameters, out of which the kernel, gamma, and random state were set to non-default values. Random state is used to promote the reproducibility of the model, whereas the max depth influences overfitting tendency and is thus included in the training process.

3.8 Computational packages selection

We have used the Scikit-Learn library for implementing the Random Forest Classifier, the XGBoost, and the SVM algorithms. Scikit-Learn is one of the most versatile and reliable library for machine learning in Python, as it features several useful tools for machine learning and statistical modeling, such as classification, regression, clustering, and dimensionality reduction via a consistency interface in Python.

Some of the features used in this study were calculated by using the following packages:

- For Sentiment Analysis calculation (of a sentence or an interview): TextBlob library, which is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.
- For Features Extraction (Average Unique Words Frequency, Average Stop Words Frequency): NLTK (Natural Language Processing Toolkit), which is a leading platform for building Python programs that work with human language data. It provides easy-to-use interfaces for over 50 corpus and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, semantic reasoning and wrappers for industrial-strength NLP libraries.

4 Results

As mentioned in the Experiment Design section, for this case study we've tested some of the most adopted machine learning models in the scientific literature with different sets of features and sets of parameters to establish how reliable are these models in the detection of depression disorders. In this sense we have adopted 3 NLP Classifiers with different set of features that were based on sentiment analysis and other natural language processing techniques.

Our results suggest that Random Forest and XGBoost are the best classifiers for the task since they have presented similar accuracy levels, exceeding 80%. Although the SVM algorithm have returned poor results in comparison with the other two models, we were able to replicate the results found in the comparable scientific literature. All the tested models results in accuracy levels beyond our baseline approach. More details about the result for each model such as features selection and parameters tuning can be found in the following sub-sections.

4.1 Baseline approach results

After arbitrarily setting our data set with a prediction value of 1, we verified the error rate (in this case the accuracy level) of approximately 65%.

Table 3. Random forest model details.

Aspect	Description
Features	Initial group of 17, 5 chosen for each classification
Parameters	Fixed
Number of Models	6,188
Best Accuracy	83.8%

https://doi.org/10.1371/journal.pone.0322299.t003

Table 4. Confusion matrix.

Class	Precision	Recall	F1-Score	Support
0	0.87	0.93	0.90	29
1	0.71	0.56	0.63	9
Accuracy			0.84	38
Macro Avg	0.79	0.74	0.76	38
Weighted Avg	0.83	0.84	0.83	38

https://doi.org/10.1371/journal.pone.0322299.t004

Table 5. Random forest - top model features with best results.

Rank	Features	Accuracy
1	speech_speed, avg_characters, avg_nouns, what's your dream job, do you consider yourself an introvert	83.8%
2	speech_speed, avg_characters, avg_nouns, what's your dream job, how are you at controlling your temper	83.8%
3	speech_speed, avg_characters, avg_nouns, what's your dream job, when was the last time you argued with someone and what was it about	83.8%
4	speech_speed, avg_characters, avg_nouns, do you consider yourself an introvert, how are you at controlling your temper	83.8%
5	speech_speed, avg_characters, avg_nouns, what's your dream job, how are you at controlling your temper	83.8%

https://doi.org/10.1371/journal.pone.0322299.t005

4.2 Random forest model results

In our model classification process, we started with an initial group of 17 features. For each classification, we selected 5 distinct features from this group. With parameters kept fixed for the classifier, this approach resulted in the creation of 6,188 unique models. From these, the top-performing models achieved a notable accuracy of approximately 83.8%. A comprehensive summary of these findings can be found in Tables 3 and 4.

The confusion matrix presented in Table 3 refers to the parameter settings shown in Table 2 and uses the feature set [speech speed, avg characters, avg nouns, how are you at controlling your temper, when was the last time you argued with someone]. However, several combinations of features yielded similar accuracy results (also satisfying the F1-Score values) as shown in Table 5

4.3 XGBoost model results

During the model training process, we experimented with varying numbers of estimators. Specifically, the "n estimators" took on values of 100, 300, and 500. By selecting 4 out of the 17 available features, we generated 2,380 unique feature sets. These were then paired with each of the 3 estimator configurations, resulting in a total of 7,140 classifiers for our model's training. Notably, the peak accuracy achieved remained consistent at around 83.8%. For a comprehensive breakdown of the features and parameters yielding the best outcomes, refer to Tables 6 and 7.

Table 6. XGBoost model details.

Aspect	Description
Features	Initial group of 17, 4 chosen for each classification
Parameters	Different "n estimators": 100, 300, 500
Number of Models	7,140
Best Accuracy	81.8%

https://doi.org/10.1371/journal.pone.0322299.t006

Table 7. Confusion matrix for features: ['speech_speed', 'avg_characters', 'adj_freq', 'how are you at controlling your temper'] with n_estimators = 300.

Class	Precision	Recall	F1-Score	Support
0	0.87	0.90	0.88	29
1	0.62	0.56	0.59	9
Accuracy			0.82	38
Macro Avg	0.75	0.73	0.73	38
Weighted Avg	0.81	0.82	0.81	38

https://doi.org/10.1371/journal.pone.0322299.t007

Table 8. XGBoost - top model features and parameters with best results for XGBoost.

Rank	Features	Parameters	Accuracy
1	speech speed, avg characters, adj freq, how are you at controlling your temper	use_label_encoder: False, eval_metric: auc, learning_rate: 0.01, max_depth: 12, n_estimators: 300	81.8%
2	speech speed, avg characters, adj freq, how are you at controlling your temper	use_label_encoder: False, eval_metric: auc, learning_rate: 0.01, max_depth: 12, n_estimators: 300	81.8%
3	avg response time, speech speed, avg nouns, how are you at controlling your temper	use_label_encoder: False, eval_metric: auc, learning_rate: 0.01, max_depth: 12, n_estimators: 100	81.8%
4	speech speed, avg characters, avg nouns, adj freq	use_label_encoder: False, eval_metric: auc, learning_rate: 0.01, max_depth: 12, n_estimators: 300	81.8%

https://doi.org/10.1371/journal.pone.0322299.t008

The confusion matrix presented in Table 7 used a subset of the parameter settings shown in Table 6 and with the the feature set [speech speed, avg characters, adj freq,how are you at controlling your temper] and n_estimators = 300. However, several combinations of features yielded similar accuracy results (also satisfying the F1-Score values) as shown in Table 8.

4.4 SVM model results

As in the previous classification models, In our model classification process we worked with an initial group encompassing 17 features. From these, we handpicked 4 features, which culminated in 2,380 unique feature sets. For our SVM model, we focused on optimizing two parameters: Kernel and Gamma. Our testing involved 4 variations of Gamma (1, 0.1, 0.01, 0.001) paired with 2 Kernel options ('rbf' and 'linear'). As a result of this comprehensive approach, we devised 76,160 SVM models.

By synergizing every feature set with the Gamma and Kernel variations, the total models per group remained consistent. Among all models tested, the pinnacle of accuracy achieved was approximately 64.8%. The detailed configurations of the model that delivered the best performance can be found in Tables 9 and 10.

Additionally,similar to the previous models, several combinations of features produced comparable accuracy and F1-score results, as illustrated in Table 11.

Table 9. SVM model details.

Aspect	Description
Features	Initial group of 17, 4 chosen for each classification
Parameters	Gamma: {1, 0.1, 0.01, 0.001}, C: {1, 10, 100, 1000}, Kernell: {'ref', 'linear'}
Number of Models	76,160
Best Accuracy	68%

https://doi.org/10.1371/journal.pone.0322299.t009

Table 10. Confusion Matrix for Features: ['avg_response_time', 'speech_speed', 'avg_characters', 'adj_freq'] with parameters: kernel = rbf, gamma = 0.1, C = 10.

Class	Precision	Recall	F1-Score	Support
0	0.87	0.69	0.77	29
1	0.40	0.67	0.50	9
Accuracy			0.68	38
Macro Avg	0.63	0.68	0.63	38
Weighted Avg	0.76	0.68	0.71	38

https://doi.org/10.1371/journal.pone.0322299.t010

Table 11. SVM - Example of top models with 4 features and best F1-score.

Features	Accuracy (%)	F1-Score (%)
avg_response_time, speech_speed, avg_characters, adj_freq	68.42	50.00
avg_response_time, speech_speed, avg_characters, what's your dream job	68.42	50.00
avg_response_time, speech_speed, avg_characters, do you consider yourself an introvert	68.42	50.00
avg_response_time, speech_speed, avg_characters, what do you do to relax	68.42	50.00
avg_response_time, speech_speed, avg_characters, how are you at controlling your temper	68.42	50.00
avg_response_time, speech_speed, avg_characters, when was the last time you argued with someone and what was it about	68.42	50.00
avg_response_time, speech_speed, avg_characters, how close are you to your family	68.42	50.00

https://doi.org/10.1371/journal.pone.0322299.t011

Table 12. Confusion Matrix for Features: ['avg_nouns', 'do you consider yourself an introvert', 'is there anything you regret', 'what's one of your most memorable experiences', 'when was the last time you felt really happy'] with parameters: kernel = rbf, gamma = 0.1, C = 10.

Class	Precision	Recall	F1-Score	Support
0	0.76	1.00	0.87	29
1	0.00	0.00	0.00	9
Accuracy			0.76	38
Macro Avg	0.38	0.50	0.43	38
Weighted Avg	0.58	0.76	0.66	38

https://doi.org/10.1371/journal.pone.0322299.t012

Finally, although the accuracy levels did not surpass the approximately 70% reported in [56], it is important to note that we identified combinations with significantly higher accuracy, as shown in Table 12. However, since these results did not meet our F1 score criteria, they were excluded from this comparison.

5 Discussion

5.1 General insights from the case study

Our results are not only better than the baseline approach in terms of accuracy but also outperformed the results obtained in the scientific literature. As a benchmark for this study we

used [56] which is an article that tested different NLP techniques for Depression detection based on the same data set used in our study.

More specifically, the Random Forest Model and XGBoost Model results represented a significant improvement in terms of accuracy (approximately 84%) in comparison with the models used by the authors in [56] (SVM and CNN with approximately 70% and 72% of accuracy, respectively).

Regarding the data cleaning e preprocessing in our case study we made text lower case and removed basic punctuation (e.g., ",", ".", "[", "]"), and also all the text content that was not actually present in the conversation, but rather indicated what kind of comment was being made by the bot. Moreover, 3 meetings were removed from the dataset due to structural problems (did not have comments –any or most – from the bot). Additionally, because the sentiment of the relevant questions had a central role among our selected features, we have realized that there was no group of questions (or single question) included in every single transcript (i.e., interview). Thus, we have used the questions that could be found in most of the interviews. In this sense, questions that were not in all meetings were assumed to yield an answer of sentiment 0 for each of the meetings in which they did not appear.

5.2 Insights from the selection of the number of features in models

Our decision to seek an optimal subset of features was driven by the realization that simply increasing the number of features does not necessarily lead to improved model performance. In some of the worst scenarios observed, using all available features led to a significant drop in performance. For instance, using all 17 features, we achieved an accuracy of 79%, in contrast to the 84% obtained with a subset of just 5 features in the Random Forest Classifier. Similarly, with the XGBoost model, the accuracy was 79% using all 17 features, while a subset of 4 features resulted in 82% accuracy. In the case of the SVM model, the use of a subset of 4 features yielded an accuracy of 68%, compared to 66% when using all features. These results suggest that some features may actually degrade the accuracy and F1-score of the models.

5.3 Insights from the sentiment features in top-performing models

When examining the top-performing models, RF and XGB, we observed a recurring prominence of features such as Average Response Time, Speech Speed, and Average Number of Characters. Intriguingly, within each group of four features, only one was tied to a sentiment score. However, this score wasn't reflective of the overall meeting sentiment but was specific to an answer to a certain question. Notably, the particular questions associated with this sentiment score varied across the best-performing models.

This leads us to several insights regarding the sentiment feature:

- The sentiment feature, linked to the response of a specific question, consistently occupies just one slot in every four-feature group.
- The exact question associated with the sentiment feature is not uniform across the top models.

From this, we can deduce that the sentiment score, specific to an answer, may not consistently influence the model's outcomes. This inconsistency hints that the sentiment attached to a particular question's response might not consistently sway the model's effectiveness. Yet, the recurring appearance of some form of sentiment feature in top models indicates that sentiment does play a part, though its influence may not be uniform. This irregularity could point

towards potential inaccuracies in how sentiment is gauged. If a sentiment score tied to a specific question was a pivotal determinant, we would expect its consistent presence across all top models. To ascertain the relevance of the sentiment feature to the classifier's results, a comprehensive feature importance analysis would be essential.

5.4 Insights from the interplay between dataset bias and imbalance in PTSD-focused studies

In our study centered on patients with PTSD, there's an inherent assumption that the dataset may lean towards a positive depression diagnosis, owing to the known comorbidity between PTSD and depression. However, it's noteworthy that our dataset was imbalanced, with only 56 of the 188 interviews being from individuals diagnosed with depression. This poses an intriguing question: might the imbalanced nature of the dataset counterbalance or mitigate the anticipated bias introduced by focusing on PTSD? Therefore, unlike [56], we relied on accuracy as a key evaluation metric for comparison, but only from the combinations with the highest F1-Score.

Still, a deeper exploration and analysis are needed to fully understand the interplay between dataset bias and imbalance in this context.

5.5 Insights related to a possible evolution from a case study to a depression detection system

Our study does not address critical issues such as the reliability of models in terms of accuracy, transparency, interpretability, and ethical impact. In a clinical setting, it is essential not only that models are accurate but also that healthcare professionals can understand and trust the decisions made by these models [62]. Similarly, as these models evolve into a comprehensive Depression Detection System, a thorough analysis of data protection measures against unauthorized access becomes crucial.

6 Limitations and threats to validity

The primary limitations of our study arise from generalization challenges associated with the inherent characteristics of our dataset. The dataset used in this study suffers from an unbalanced class distribution. Additionally, it is biased due to its origin from a psychiatric population (i.e., participants with PTSD), which introduces concerns related to fairness, particularly since the dataset includes participants of both genders, yet these gender differences were not explicitly addressed. Despite these limitations, the results are promising when evaluating various alternatives, including models, data cleaning and preprocessing techniques, selected features, and parameters. Finally, these limitations are addressed in the Future Work section.

7 Conclusions and future work

7.1 Conclusions

This initial study focuses on developing a Depression Detection Model using Sentiment Analysis and other NLP techniques specifically for patients diagnosed with PTSD. In our research, we compared various classifiers, feature sets, and other factors to determine the most effective approach.

We have trained our models with a dataset built from the transcripts from 188 sessions of clinical interviews which correspond to questionnaire responses and audio and video recordings. The interviews were conducted by autonomous agents (bot). Our approach enabled us to

achieve better results in terms of accuracy then we had in the comparable literature based on the same data set. The main difference between our case study and the comparable literature relied precisely on the model choices, features selection and data cleaning and processing of our data set.

After determining the best techniques, models and features, we tested thousands (sometimes tens of thousands) of combinations of features and parameters to achieve optimal results for each model and technique. This exhaustive analysis not only enhanced the accuracy and reliability of our chosen models but also provided a deeper understanding of the intricate relationships between various features and parameters. This insight paves the way for more targeted and efficient implementations in subsequent studies.

In summary, we identified compelling evidence indicating numerous opportunities for advancement in the future development of a depression diagnostic system utilizing NLP techniques. These potential improvements are multifaceted. They encompass the exploration of a wider range of model choices, rigorous data cleaning and pre-rocessing methodologies, and, most prominently, refined strategies in feature selection and intricate feature engineering.

7.2 Future work

The findings of our study chart the way for several avenues of future research, primarily centered around expanding our methodology, refining feature selection, enhancing data preprocessing, and bolstering model generalization.

In terms of model selection, our comparative analysis could be broadened to include other ML classifiers such as Decision Trees, Convolutional Neural Networks (CNN), and BERT (Bidirectional Encoder Representations from Transformers). Given the state-of-the-art performance of BERT and its derivatives (e.g., RoBERTa, DistilBERT) across various NLP tasks, we propose exploring CNN and BERT-based models for feature generation, particularly by utilizing vector space representations for textual data. Further optimization studies for kernel algorithms might enhance the performance of the SVM model. Moreover, adopting Transformer models like BERT and leveraging Large Language Models (LLM) can refine sentiment score calculations, enhancing our classifiers' feature selection. This progression should culminate in a detailed feature importance analysis.

In terms of feature selection enhancement, lexicon-based approaches using dictionaries specific to medical conditions can be invaluable. Generating features such as word frequency, sentiment, and more can be beneficial. Additionally, integrating TF-IDF for pinpointing significant transcript words and incorporating bigrams can enrich contextual understanding and sentiment score assignment. An intriguing perspective we encountered centers on assigning variable weights to negative words based on context, achieved through combining lexicon dictionaries and context-weighted sentiment analysis.

Additionally, to enhance feature selection in datasets derived from DAIC-WOZ or similar datasets, specific techniques can be adapted as demonstrated in various articles. For instance, in [33], the Ant-Lion Optimization (ALO) algorithm could be applied to identify relevant NLP-based features like speech speed, sentiment, and average noun usage, which would allow for the reduction of redundant or less significant features, improving classifier accuracy. Similarly, in [32], the hybrid approach combining Emperor Penguin Optimization (EPO) and Bacterial Foraging Optimization (BFO) can be tailored to optimize feature selection from textual and acoustic data, enhancing the early detection of conditions such as depression. In [34] the authors offer an effective method that could be leveraged to refine the selection of NLP features, ensuring a more robust model generalization. Finally, [31] provides insights into

using a combination of feature selection methods that could be adapted to optimize sentiment analysis features, leading to better diagnostic support in mental health assessments.

Exploring new features such as the frequency of the word "I", use of possessive pronouns, past tense verb occurrences, and readability scores can add depth to our analysis. We also advocate for refined preprocessing methods, such as leveraging the original NLTK set for stop word removal and incorporating lemmatization, a staple in NLP and machine learning.

Addressing dataset bias and imbalance remains pivotal. Potential research angles include the use of transfer learning, which presents advantages in scenarios with limited labeled data, reducing the risk of overfitting and enhancing performance on target tasks. Also, we can assess if the dataset imbalance can mitigate biases introduced by emphasizing PTSD sufferers, given their established association with depression. It is also relevant to implement rebalancing techniques that can enhance model generalization, and explore larger, more balanced datasets to prevent overfitting and ensure robust model generalization. For instance, downsampling could be employed to mitigate the unbalanced class distribution by reducing the overrepresentation of the majority class, thereby helping the model to treat both classes with equal importance and improve its generalization to new data. Additionally, stratification could be used to address fairness concerns by ensuring that the gender distribution (and other relevant factors) is proportionally represented in all subsets of the data, including training, validation, and test sets. This would help to prevent the model from developing biases towards a particular gender and ensure more equitable performance across different groups.

In future work, we plan to refine our model evaluation process by introducing a development set specifically for classifier selection. This approach will allow us to optimize models independently from the test set, ensuring a more robust and unbiased assessment of their performance. By evaluating the selected classifiers on a separate test set, we aim to provide a clearer and more accurate representation of the models' generalization capabilities.

Finally, after exploring all potential expansions of this case study, a logical progression would be the development of a comprehensive Depression Detection System. In this scenario, it would be crucial to evaluate the reliability of the models and ensure the privacy of sensitive data. In this context, the q-ROF2TL-FWZIC method should be considered to assess various aspects of the model, such as accuracy, transparency, interpretability, and ethical impact, by assigning appropriate weights to these criteria according to their importance [62]. Additionally, it is essential to explore medical information security techniques.

Author contributions

Conceptualization: Giuliano Lorenzoni.

Data curation: Cristina Tavares.

Formal analysis: Giuliano Lorenzoni.

Investigation: Giuliano Lorenzoni.

Methodology: Giuliano Lorenzoni, Paulo Alencar, Donald Cowan.

Project administration: Giuliano Lorenzoni, Nathalia Nascimento, Paulo Alencar, Donald Cowan.

Resources: Cristina Tavares, Nathalia Nascimento.

Software: Giuliano Lorenzoni, Paulo Alencar, Donald Cowan.

Supervision: Nathalia Nascimento, Paulo Alencar, Donald Cowan.

Validation: Giuliano Lorenzoni, Nathalia Nascimento, Paulo Alencar, Donald Cowan.

Visualization: Giuliano Lorenzoni, Nathalia Nascimento.

Writing – original draft: Giuliano Lorenzoni, Paulo Alencar.

Writing – review & editing: Giuliano Lorenzoni, Cristina Tavares, Nathalia Nascimento, Paulo Alencar, Donald Cowan.

References

1. DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, et al. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS). Richland, SC. IEEE; 2014. p. 1061–8.
2. Saqib K, Khan AF, Butt ZA. Machine learning methods for predicting postpartum depression: scoping review. JMIR Ment Health. 2021;8(11):e29838. https://doi.org/10.2196/29838 PMID: 34822337
3. Lee Y, Ragguett R-M, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord. 2018;241:519–32. https://doi.org/10.1016/j.jad.2018.08.073 PMID: 30153635
4. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform. 2020;8(3):e17984. https://doi.org/10.2196/17984 PMID: 32229465
5. Ganguli R, Mehta A, Sen S. A survey on machine learning methodologies in social network analysis. In: 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). IEEE; 2020. p. 484–9.
6. 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). IEEE; 2021. p. 718–23. https://doi.org/10.1109/icais50930.2021.9395943
7. Nandanwar H, Nallamolu S. Depression prediction on Twitter using machine learning algorithms. In: 2nd Global Conference for Advancement in Technology (GCAT). IEEE; 2021. p. 1–7.
8. Yuki JQ, Sakib MdMQ, Zamal Z, Efel SH, Khan MA. Detecting depression from human conversations. In: Proceedings of the 8th International Conference on Computer and Communications Management. 2020. p. 14–8. https://doi.org/10.1145/3411174.3411187
9. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in reddit social media forum. IEEE Access. 2019;7:44883–93. https://doi.org/10.1109/access.2019.2909180
10. Azam F, Agro M, Sami M, Abro MH, Dewani A. Identifying depression among twitter users using sentiment analysis. In: 2021 International Conference on Artificial Intelligence (ICAI). IEEE; 2021. p. 44–9. https://doi.org/10.1109/icai52203.2021.9445271
11. Hossain MdT, Rahman Talukder MdA, Jahan N. Social networking sites data analysis using NLP and ML to predict depression. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE; 2021. p. 1–5. https://doi.org/10.1109/icccnt51525.2021.9579916
12. Skaik R, Inkpen D. Using twitter social media for depression detection in the Canadian population. In: Proceedings of the 3rd Artificial Intelligence and Cloud Computing Conference; 2020. p. 109–14.
13. Alghamdi NS, Hosni Mahmoud HA, Abraham A, Alanazi SA, Garcia-Hernandez L. Predicting depression symptoms in an Arabic psychological forum. IEEE Access. 2020;8:57317–34. https://doi.org/10.1109/access.2020.2981834
14. Rastogi N, Keshtkar F, Miah MS. A multi-modal human robot interaction framework based on cognitive behavioral therapy model. In: Proceedings of the Workshop on Human-Habitat for Health (H3): Human-Habitat Multimodal Interaction for Promoting Health and Well-Being in the Internet of Things Era; 2018. p. 1–6.
15. Zhang P, Wu M, Dinkel H, Yu K. Depa: Self-supervised audio embedding for depression detection. In: Proceedings of the 29th International Conference on Multimedia. ACM; 2021. p. 135–43.
16. Maniar S, Patil K, Rao B, Shankarmani R. Depression detection from Tweets along with clinical tests. In: International Conference on Intelligent Technologies (CONIT). IEEE; 2021. p. 1–6.
17. Zogan H, Razzak I, Jameel S, Xu G. Depressionnet: a novel summarization boosted deep framework for depression detection on social media. arXiv preprint. 2021. https://arxiv.org/abs/2105.10878

18. Jain MP, Sribash Dasmohapatra S, Correia S. Mental Health State Detection using Open CV and sentimental analysis. In: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE; 2020. p. 465–70. https://doi.org/10.1109/iciss49785.2020.9315984
19. Trotzek M, Koitka S, Friedrich CM. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. In: Transactions on Knowledge and Data Engineering (TKDE), vol. 32, no. 3. IEEE; 2018. p. 588–601.
20. Uddin MZ, Dysthe KK, Følstad A, Brandtzaeg PB. Deep learning for prediction of depressive symptoms in a large textual dataset. Neural Comput Appl. 2021;34(1):721–44. https://doi.org/10.1007/s00521-021-06426-4
21. Wongkoblap A, Vadillo MA, Curcin V. Deep learning with anaphora resolution for the detection of tweeters with depression: algorithm development and validation study. JMIR Ment Health. 2021;8(8):e19824. https://doi.org/10.2196/19824 PMID: 34383688
22. Alabdulkreem E. Prediction of depressed Arab women using their tweets. J Decis Syst. 2020;30(2–3):102–17. https://doi.org/10.1080/12460125.2020.1859745
23. Wang Y, Wang Z, Li C, Zhang Y, Wang H. A multimodal feature fusion-based method for individual depression detection on Sina Weibo. In: 39th International Performance Computing and Communications Conference (IPCCC). IEEE; 2020. p. 1–8.
24. Mulay A, Dhekne A, Wani R, Kadam S, Deshpande P, Deshpande P. Automatic depression level detection through visual input. In: 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). IEEE; 2020. p. 19–22. https://doi.org/10.1109/worlds450073.2020.9210301
25. 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). IEEE; 2020. p. 275–80. https://doi.org/10.1109/bibe50027.2020.00052
26. Sanyal H, Shukla S, Agrawal R. Study of depression detection using deep learning. In: 2021 IEEE International Conference on Consumer Electronics (ICCE). IEEE; 2021. p. 1–5. https://doi.org/10.1109/icce50685.2021.9427624
27. A. Musleh D, A. Alkhales T, A. Almakki R, E. Alnajim S, K. Almarshad S, S. Alhasaniah R, et al. Twitter Arabic sentiment analysis to detect depression using machine learning. Comput Mater Continua. 2022;71(2):3463–77. https://doi.org/10.32604/cmc.2022.022508
28. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during COVID-19: observational study. J Med Internet Res. 2020;22(10):e22635. https://doi.org/10.2196/22635 PMID: 32936777
29. Nguyen T, Phung D, Dao B, Venkatesh S, Berk M. Affective and content analysis of online depression communities. IEEE Trans Affective Comput. 2014;5(3):217–26. https://doi.org/10.1109/taffc.2014.2315623
30. Zaghouani W. A large-scale social media corpus for the detection of youth depression (project note). Procedia Comput Sci. 2018;142:347–51. https://doi.org/10.1016/j.procs.2018.10.483
31. Singh LK, Khanna M, Singh R. An enhanced soft-computing based strategy for efficient feature selection for timely breast cancer prediction: Wisconsin Diagnostic Breast Cancer dataset case. Multimedia Tools and Applications. 2024. p. 1–66.
32. Singh LK, Khanna M, Singh R. Feature subset selection through nature inspired computing for efficient glaucoma classification from fundus images. Multimedia Tools and Applications. 2024. p. 1–72.
33. Khanna M, Singh LK, Shrivastava K, Singh R. An enhanced and efficient approach for feature selection for chronic human disease prediction: a breast cancer study. Heliyon. 2024;10(5):e26799. https://doi.org/10.1016/j.heliyon.2024.e26799 PMID: 38463826
34. Singh LK, Garg H. A novel approach for human diseases prediction using nature inspired computing & machine learning approach. Multimed Tools Appl. 2023;83(6):17773–809. https://doi.org/10.1007/s11042-023-16236-6
35. Jagtap N, Shukla H, Shinde V, Desai S, Kulkarni V. Use of ensemble machine learning to detect depression in social media posts. In: Second International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE; 2021. p. 1396–400.
36. Hassan SB, Zakia U. Recognizing suicidal intent in depressed population using NLP: a pilot study. In: 11th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE; 2020. p. 121–8.
37. Haque F, Nur RU, Al Jahan S, Mahmud Z, Shah FM. A transformer-based approach to detect suicidal ideation using pre-trained language models. In: 23rd International Conference on Computer and Information Technology (ICCIT). IEEE; 2020. p. 1–5.

38. Obagbuwa IC, Danster S, Chibaya OC. Supervised machine learning models for depression sentiment analysis. Front Artif Intell. 2023;6:1230649. https://doi.org/10.3389/frai.2023.1230649 PMID: 37538396
39. Liu D, Feng XL, Ahmed F, Shahid M, Guo J. Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Ment Health. 2022;9(3):e27244. https://doi.org/10.2196/27244 PMID: 35230252
40. Bhavya S, Dmello RC, Nayak A, Bangera SS. Speech emotion analysis using ML for depression recognition: a review; 2022.
41. Sawalha J, Yousefnezhad M, Shah Z, Brown MRG, Greenshaw AJ, Greiner R. Detecting presence of PTSD using sentiment analysis from text data. Front Psychiatry. 2022;12:811392. https://doi.org/10.3389/fpsyt.2021.811392 PMID: 35178000
42. Chaves A, Kesiku C, Garcia-Zapirain B. Automatic text summarization of biomedical text data: a systematic review. Information. 2022;13(8):393. https://doi.org/10.3390/info13080393
43. Squires M, Tao X, Elangovan S, Gururajan R, Zhou X, Acharya UR, et al. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. Brain Inform. 2023;10(1):10. https://doi.org/10.1186/s40708-023-00188-6 PMID: 37093301
44. Cummins N, Baird A, Schuller BW. Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. Methods. 2018;151:41–54. https://doi.org/10.1016/j.ymeth.2018.07.007 PMID: 30099083
45. Krishna S, Anju J. Different approaches in depression analysis: a review. In: 2020 International Conference on Computational Performance Evaluation (ComPE). 2020. p. 407–14. https://doi.org/10.1109/compe49325.2020.9200001
46. Nguyen TT, Pham VH-Q, Le D-T, Vu X-S, Deligianni F, Nguyen HD. Multimodal machine learning for mental disorder detection:a scoping review. Procedia Comput Sci. 2023;225:1458–67. https://doi.org/10.1016/j.procs.2023.10.134
47. Nouman M, Khoo SY, Mahmud MAP, Kouzani AZ. Recent advances in contactless sensing technologies for mental health monitoring. IEEE Internet Things J. 2022;9(1):274–97. https://doi.org/10.1109/jiot.2021.3097801
48. da Silva AC, Santana RC, de Lima TH, Teodoro ML, Song MA, Zárate LE, et al. A review of computational methods and databases used in depression studies. HEALTHINF. 2022. p. 413–20.
49. Yan W-J, Ruan Q-N, Jiang K. Challenges for artificial intelligence in recognizing mental disorders. Diagnostics (Basel). 2022;13(1):2. https://doi.org/10.3390/diagnostics13010002 PMID: 36611294
50. Song S, Shen L, Valstar M. Human behaviour-based automatic depression analysis using hand-crafted and deep spectral features. In: 13th International Conference on Automatic Face and Gesture Recognition (FG). IEEE; 2018. p. 158–65.
51. Janardhan N, Kumaresh N. Improving depression prediction accuracy using fisher score-based feature selection and dynamic ensemble selection approach based on acoustic features of speech. TS. 2022;39(1):87–107. https://doi.org/10.18280/ts.390109
52. Mao K, Wu Y, Chen J. A systematic review on automated clinical depression diagnosis. NPJ Ment Health Res. 2023;2(1):20. https://doi.org/10.1038/s44184-023-00040-z PMID: 38609509
53. Aleem S, Huda N ul, Amin R, Khalid S, Alshamrani SS, Alshehri A. Machine learning algorithms for depression: diagnosis, insights, and research directions. Electronics. 2022;11(7):1111. https://doi.org/10.3390/electronics11071111
54. Ndaba S. Review of class imbalance dataset handling techniques for depression prediction and detection. 2023. Available at SSRN: 4387416.
55. Yang L, Jiang D, He L, Pei E, Oveneke MC, Sahli H. Decision tree based depression classification from audio video and language information. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. 2016. p. 89–96. https://doi.org/10.1145/2988257.2988269
56. Yang C, Lai X, Hu Z, Liu Y, Shen P. Depression tendency screening use text based emotional analysis technique. J Phys: Conf Ser. 2019;1237(3):032035. https://doi.org/10.1088/1742-6596/1237/3/032035
57. Amershi S, Begel A, Bird C, DeLine R, Gall H, Kamar E, et al . Software engineering for machine learning: a case study. In: 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). IEEE/ACM; 2019. p. 291–300.
58. ScikitLearn. Statistical learning: the setting and the estimator object in scikit-learn; 2007.
59. Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. In: LREC. 2014. p. 3123–8.
60. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord. 2009;114(1–3):163–73. https://doi.org/10.1016/j.jad.2008.06.026 PMID: 18752852

61. Tavares C, Nascimento N, Alencar P, Cowan D. Adaptive method for machine learning model selection in data science projects. In: International Conference on Big Data (Big Data). IEEE; 2022. p. 2682–8.

62. Alsalem MA, Alamoodi AH, Albahri OS, Albahri AS, Martínez L, Yera R, et al. Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach. Exp Syst Appl. 2024;246:123066. https://doi.org/10.1016/j.eswa.2023.123066