Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection

Bayode Ogunleye *, Hemlata Sharma, Olamilekan Shobayo *

Posted Date: 16 July 2024

doi: 10.20944/preprints202407.1325.v1

Keywords: Depression detection; large language models; machine learning, mental health; natural language processing, public health

 doi:10.20944/preprints202407.1325.v1

Article

Sentiment Informed Sentence BERT-Ensemble Algorithm for Depression Detection

Bayode Ogunleye [1,*], Hemlata Sharma [2] and Olamilekan Shobayo [2,*]

[1] Department of Computing & Mathematics, University of Brighton, Brighton BN2 4GJ, United Kingdom
[2] Department of Computing, Sheffield Hallam University, Sheffield, S1 2NU, United Kingdom; h.sharma@shu.ac.uk
* Correspondence: b.ogunleye@brighton.ac.uk (B.O.); o.shobayo@shu.ac.uk (O.S.)

Abstract: The world health organisation (WHO) revealed approximately 280 million people in the world suffer from depression. Yet, existing studies on early-stage depression detection using machine learning (ML) techniques are limited. Prior studies have applied a single stand-alone algorithm which are unable to deal with data complexities, prone to overfitting and limited in generalisation. To this end, our paper examined the performance of several ML algorithms for early-stage depression detection using two benchmark social media datasets (D1 and D2). More specifically, we incorporated sentiment indicator to improve our model performance. Our experimental results showed that sentence bidirectional encoder representations from transformers (SBERT) numerical vectors fitted into stacking ensemble model achieved comparable F1 scores of 69% in dataset (D1) and 76% in dataset (D2). Our findings suggest that utilising sentiment indicators as additional feature for depression detection yields an improved model performance and thus, we recommend the development of depressive term corpus for future work.

Keywords: depression detection; large language models; machine learning; mental health; natural language processing; public health

1. Introduction

There is a growing concern as regards the deteriorating mental health of people [1,2]. WHO [3] reported a growing drop in the mental health of people as individuals diagnosed of depression has increased. In recent times, depression has garnered global attention due to its widespread prevalence and the significant risk it poses, particularly in terms of suicide [1,3,4]. Depression is one of the leading causes of disability globally [5]. WHO [3] reported approximately 280 million people in the world suffer from depression. Their findings showed more women are affected than men and 5% of adults suffer from depression. Literature evidences the symptoms of depression as unhappiness, loneliness, forgetfulness, low self-esteem, thoughts of self-harm, disrupted sleep, loss of energy, changes in appetite, anxiety, reduced concentration, indecision, feelings of worthlessness, lack of interest, guilt, or hopelessness [2,6]. Previous studies showed that primary causes of depression are financial issues, workplace issues, family issues, and academic performance [7]. There are traditional approaches to depression detection which requires psychometric questions. Several depression questionnaires had been developed. For example, Beck's Depression Inventory [8], Center for Epidemiologic Studies Depression Scale [9], Children's Depression Inventory [10], Mood and Feelings Questionnaire [11], Patient Health Questionnaire-9 [12], and Revised Children's Anxiety and Depression Scale [13]. However, this approach is labour intensive, time consuming and limited to pre-defined attributes. In addition, research evidenced that people with depression may not know or acknowledge they have the illness and thus influences the survey outcome for early-stage depression screening [14,15].

 doi:10.20944/preprints202407.1325.v1

People obtain information or express their feelings freely on social media [1,16,17]. Thus, user generated contents (UGC) are considered a reliable data source for developing automated depression detection model [18,19]. Interestingly, the rapid development of machine learning approaches has made significant contributions to the development of depression detection algorithms [20]. Thus, this study will focus on the use of machine learning (ML) techniques. Several machine learning algorithms were implemented for depression detection. For example, K-nearest neighbours [21], support vector machine [22], decision tree [23], XGBoost [24], multinomial Naïve Bayes [25], convolutional neural network [26], recurrent neural network [27], long short-term memory [28], Bidirectional long short-term memory [29] and RoBERTa [30]. Unfortunately, a major concern of existing algorithms is the generalisation issue [4,19,31,32]. Cai et al. [4] collected data from Chinese social network platform Sina Weibo. The dataset consists of 742,430 depressed tweets and 729,409 non-depressed selected from 3,711 depressed users and 19,526 non-depressed users. Thus acknowledged the limitation of their study is generalisation issue. This is because depressed users tend to post less tweets than non-depressed users on Twitter, while users on Sina Weibo behave just the opposite. Furthermore, in the context of depression detection, using traditional classification methods may not succeed, due to the difficulty in extracting discriminative features from texts on social media [1,33]. Lastly, majority of studies found have utilised a single stand-alone algorithm which are unable to deal with data complexities, prone to overfitting and limited in generalisation, and thus, effectiveness of the models is unsatisfactory for real-world deployment [1]. Based on this background, our aim is to develop a well performing ML algorithm for depression detection. To this end, we propose the use of Sentence BERT-Ensemble model for depression detection. Thus, our main contributions can be summarised as follows.

1. We conduct an experimental comparison of several state-of-the-art (SOTA) ML algorithms for depression detection and discuss them from a scientific lens.
2. We demonstrate the use of Sentence BERT-Ensemble model to achieve SOTA results.
3. We demonstrate that sentiment analysis indicator is a useful external feature in depression detection.

Subsequent sections present the background knowledge to this study (Section 2), methodology (Section 3), results (Section 4) and Section 5 will provide conclusions and recommendations.

2. Related Works

Previous studies have proposed several statistical and ML methods for depression detection. For example, Figuerêdo et al. [1] proposed convolutional neural network (CNN) in combination with context-independent word embeddings and fusion based approaches. They used dataset collected from Reddit users (in English). In their experiment, they trained the algorithms using 30,851 depressed posts and 264,172 non-depressed. Thereafter, they tested with 18,706 depressed posts and 217,665 non-depressed and their approach achieved a precision of 76% and F1 score of 71%. Shrestha et al. [34] applied five variants of bidirectional encoder representations from transformers (BERT) with and without fine-tuned layers for depression detection. Their study focused on identifying moderate depression using typed and transcribed responses from the StudentSADD dataset. Their findings highlighted that typed responses hold more value than transcribed responses in depression screening. They showed DistilBERT with fine-tuning layers demonstrated the best performance for typed responses with accuracy of 63%. Naseem et al. [35] formulated depression detection as a multi-classification task by recategorizing the depression severity levels as minimal, mild, moderate, and severe. They showed embeddings from text graph convolutional network (TextGCN) feed into Bidirectional long short-term memory (BiLSTM) with attention layer, and ordinal classification layer outperforms support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), CNN, recurrent neural network (RNN), long short-term memory (LSTM), BiLSTM and DepressionNet across two different Reddit datasets. Their approach achieved an F1 scores of 85% (in their D1 dataset of 3553 posts) and 95% (in their D2 dataset of 77350 posts). Pérez et al. [19] made publicly available BDI-Sen dataset comprising of 4973 annotated sentences covering depressive symptoms and 41,200 control sentences for depression detection research. In their study, they showed MentalBERT

 doi:10.20944/preprints202407.1325.v1

(MBERT) outperformed BERT, BERT-mini, T5 (text-to-text transfer transformer), logistic regression (LR), and SVM as a binary classification task with an f1 score of 83%. Similarly, they performed a fine-grained depression severity level and showed MBERT outperformed other models. Monreale et al. [36] used psychometric profile of users as features to improve the performance of SVM for depression detection and thus, achieved an accuracy of 96%. Sen et al. [37] aimed to understand the effect of company and region on measuring workplace depression. Their study adopted the occupational depression inventory (ODI) scale designed for work-related depression assessment. They collected over 358,527 employee reviews from 104 prominent US companies from 2008 to 2020. In addition, collected yearly stock growth values of the 104 companies from Yahoo financial portal, and values of five social economic indicators. Thus, proposed AutoODI1 framework which uses SBERT and cosine similarity to assign composite ODI scores to the reviews. Their findings suggest US states that host companies with high ODI scores also manifested high depression rates, talent shortage, and economic deprivation. Wu et al. [38] proposed a deep neural network for the early prediction of depression risk. Their model leveraged daily mood fluctuations as a psychiatric indicator and incorporated textual and emotional attributes through knowledge distillation and achieved an AUROC of 0.9317 and AUPRC of 0.8116. Villatoro-Tello et al. [39] prepared lexicon and used them as features for the development of depression detection binary classification model. They used the patient's health questionnaire (PHQ-8) score to determine the level of patient's depression. They performed their experiment using the Distress Analysis Interview Corpus - wizard of Oz (DAIC-WOZ) dataset and the Extended Distress Analysis Interview Corpus (E-DAIC), which is an extended version of the DAIC-WOZ dataset and achieved a macro F1-score of 83% and 80% respectively. They showed their lexical availability theory approach outperformed BERT and MLP in a Clinical interview context. However, it is worth stating that their implementation did not handle the class imbalance issue which might impact generalisation. Zhang et al. [6] proposed a Span-based PHQ-aware and similarity contrastive network (SpanPHQ) for identifying depressive symptoms. Their model took into consideration both semantic contextual features and PHQ-9 descriptive features. Their approach evidenced superior performance compared to BERT and its variants. Liu et al. [40] argued that ecological momentary assessment (EMA) captures the mood better than standard questionnaires like PHQ-9. To evidence this, they analysed the EMA data and passive sensing data to generate synergy which can be deployed to enhance the performance for depression mood detection. They used their approach to enhance several ML algorithms including SVM, LR, KNN, RF and decision tree (DT). Their models were applied to the StudentLife dataset and achieved an average accuracy of 98%. Malik et al. [41] used the Hamilton rating scale to determine the level of people's depression, The survey results (1694 records) were modelled for depression detection using ML algorithms including KNN, DT and naïve bayes (NB). They showed KNN was superior with an accuracy of 92%. Authors in Gallegos Salazar et al. [42] proposed a contrast pattern-based classifier to detect depression by using a new data representation based only on emotion and sentiment analysis. Similarly, Amanat et al. [28] showed RNN-LSTM outperformed SVM, NB, CNN, and decision tress with an accuracy of 99%. Burdisso et al. [43] proposed SS3 (smoothness, significance, and sanction) framework to improve on incremental classification, support for early classification, and explainability for text classification algorithm. They compared SS3 to LR, SVM, KNN, and NB. Their approach achieved with F1 score of 61%. Trotzek et al. [44] proposed a hybrid approach for depression detection. In their study, they developed a LR classifier by incorporating readability and emotion features into user-level linguistic metadata and enhanced its performance using CNN. Adarsh et al. [45] addressed the problem of imbalance present across various age groups and demographics through the utilization of the one-shot decision approach. They used an innovative ensemble model that combines SVM and KNN techniques and achieved an accuracy of 98.05%. Guo et al. [46] constructed Chinese depression lexicon by expanding the Dalian university of technology sentiment lexicon (DUT-SL) with depression specific vocabulary (with the help of domain expertise). They proposed an ensemble method based on the newly constructed lexicon and feature fusion approach. Thus, showed their approach outperformed ML algorithms such as logistic regression, RF,

 doi:10.20944/preprints202407.1325.v1

KNN, SVM, DT, eXtreme Gradient Boosting (XGBoost) and light gradient boost machine (LightGBM). In the next section, this paper details the methodology adopted in this study.

3. Methodology

This section presents the methods applied in this study. This study conducted an extensive and thorough comparison of ML algorithms ranging from the traditional ML, deep learning, and transformer-based algorithms. In our attempt to improve generalisation and ascertain the SOTA approach, we performed several experiments using two datasets as outlined below.

I. In the first experiment, we compared traditional ML algorithms using term frequency and inverse document frequency (TF-IDF) vectorizer.
II. In the second attempt, we compared ML algorithms using contextual word embeddings such BERT and SBERT.
III. Finally, we implemented sentiment analysis and used the polarity result as an explicit feature. Thus, compared ML algorithms using the contextual word embeddings.

3.1. Proposed Approach

This paper compared several ML approaches that have shown good performances in the depression detection literature. Thus, based on our experiment, this study proposes the use of sentence BERT embedding and stacked ensemble model.

The experimental set up of our approach is shown in Figure 1. The raw data is converted into word embeddings numerical vector models such as BERT. In parallel, sentiment analysis is performed and thus, the numerical vectors are fed to the classifier (stacked ensemble model) to provide the required prediction. Subsequent sections discuss the individual components of our proposed approach for depression detection, presenting the algorithms used in each stage.

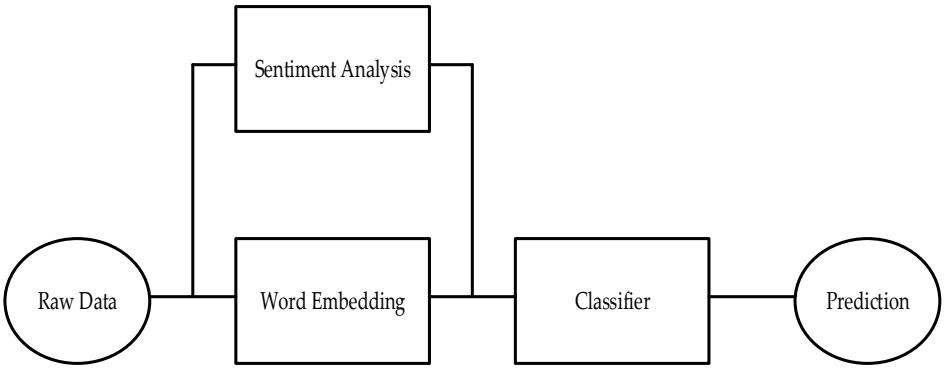

Figure 1. An illustration of our experimental setup.

3.1.1. BERT (Bidirectional Encoder Representations from Transformers)

Given a sequence of input embeddings $X = x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ the BERT model [47] consists of an encoder stack that utilizes self-attention mechanisms. The self-attention mechanism involves three sets of weight matrices Query(Q), Key(K) and Value (V). These are used to compute attention scores, and the output of the self–attention layer is obtained by applying these attention scores to the values.

The self–attention mechanism is mathematically represented as:

$$Attention\,(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Here, $Q = XW_Q$,
$K = XW_K$
$V = XW_V$
d_k is the dimensionality of the key vectors.

 doi:10.20944/preprints202407.1325.v1

BERT stacks multiple layers of these self-attention mechanisms to capture conceptual information in the input sequence. BERT uses a masked Language model (MLM) objective during pre-training, where some of the input words are masked, model is trained to predict the masked words based on the context.

3.1.2. Sentence-BERT

Sentence BERT (SBERT) [48] is a variant of BERT, designed to reduce the complexity of the computational overhead presented in BERT algorithm for regression task such as sentence-pair similarity. It has also been used for classifying sentence-pair based on the labelled class (binary or multi-class). SBERT framework with classification objective (used for this work) consist of a Siamese bi-encoder, that comprises of 2 fine-tuned BERT models each providing different output vector embeddings u and v. A third vector $|u - v|$ is also obtained, which is the element-wise absolute difference between the vector embeddings of the Siamese BERT encoder. The vectors are then concatenated and multiplied by a weight matrix W and the output is sent to a SoftMax layer to generate the probabilities of each class. This is illustrated in Figure 2.

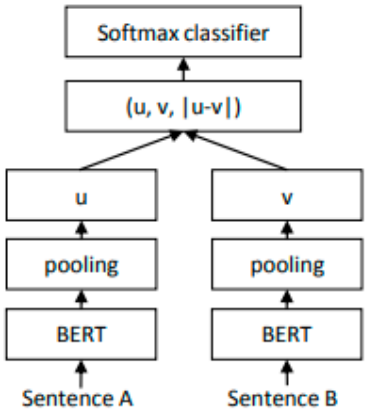

Figure 2. Classification objective framework of the SBERT model [48].

3.1.3. Stacking Ensemble Model

Ensemble techniques have experienced increase in relevance in recent years. This can be attributed to the increased performance it provides when compared to using individual models for machine learning task such as regression and classification. It was first introduced in the early 90's with the idea that individual algorithms with weak prediction could be improved to provide better predictions by combining them in some specific or methodological order [49]. The most common methods used in ensemble learning include Bagging, Boosting, and Stacking. Each of these methods varies slightly with their architecture and their selection are dependent on the nature of the ML task. For this work, we have used the stacking ensemble framework.

Stacking, also known as stacked generalisation, is an ensemble learning strategy that generates predictive outcomes of selected initial learner models known as the base models to a final classifier model, known as the meta-classifier, to obtain the final classification for the specific ML task and in most cases, improves the overall performance of the classification when compared to the individual base models and has seen it use in different classification tasks ranging from healthcare to agri-business [50–53]. As illustrated in Figure 3, stacking provides model variety, allowing all base classifiers to participate in the learning process which in turn helps reduce bias. It also provides interpretability of the learning capacity of the base models and how they influence the results of the selected meta-classifier. The diversity among the chosen base models helps the meta-classifier produce improved classification result. This is due to the different principles of generalization by the

 doi:10.20944/preprints202407.1325.v1

base models. The diversity provided by the models is obtained by using learner models of varying learning strategies, thereby introducing a level of disagreement in pattern recognition [54].

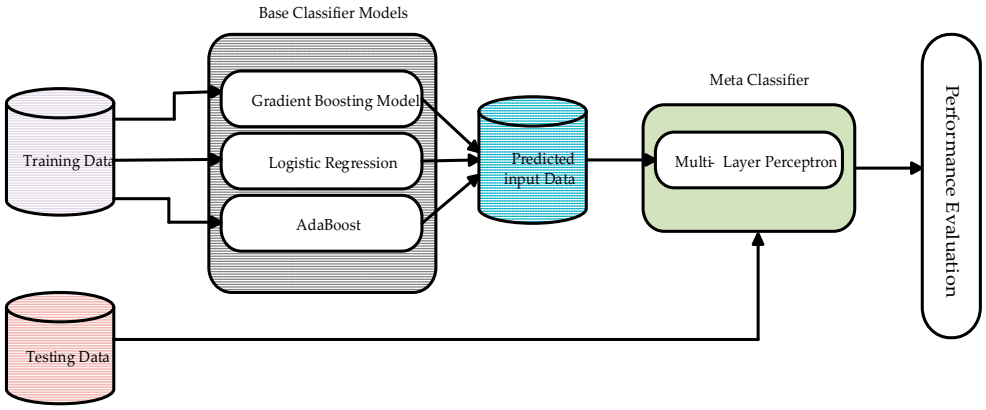

Figure 3. Illustration of stacked ensemble model.

The ensemble technique can be represented using the equation:

$$\hat{y} = f_{meta}[b_{mi}(y_i \dots \dots \dots . y_N)] \quad (2)$$

Where $\hat{y}$ represents the predicted output based on the meta classifier function f_{meta} on the output of the selected individual base models b_m. With y_i representing the predicted output of the individual base models. In subsequent section, we discuss individual ML algorithms deployed.

3.1.4. Gradient Boosting

Gradient Boosting is a powerful ensemble technique in machine learning. Unlike traditional models that learn from the data independently, boosting combines the predictions of multiple weak learners to create a single, more accurate strong learner. The key concept is to sequentially add weak learners, each one correcting the errors of its predecessors. Mathematically, the gradient boosting algorithm can be expressed as follows [55]

$$g_t(x) = E_y \left[\frac{\partial \Psi(\, y, f(x))}{\partial f(x)} | x \right]_{f(x)=\hat{f}^{t-1}(x)} \quad (3)$$

Where; $\Psi(\mathrm{y}, \mathrm{f}(\mathrm{x}))$ is the loss function measuring the difference between the actual value of y and the predicted value $f(x)$.

$\hat{f}^{t-1}(x)$ is the prediction from the model at iteration $t-1$.

$g_t(x)$ represents the pseudo-residuals, the negative gradients of the loss function with respect to the current model predictions.

In each iteration, a new weak learner $h^t(x)$ is trained to predict these pseudo-residuals. The model is then updated by adding the new learner, scaled by a learning rate η: $f^t(\mathrm{x}) = \mathrm{f}^{t-1}(\mathrm{x}) + \eta \mathrm{ht}(\mathrm{x})$. This process continues for a predefined number of iterations, resulting in a strong predictive model that minimizes the loss function effectively.

3.1.5. Logistic Regression

LR is a statistical method for modelling the probability of a binary outcome based on one or more predictor variables. It is widely used for classification problems where the dependent variable is categorical [56]

$$P(X) = \frac{e^{(b_0 + b_1^* X)}}{1 + e^{(b_0 + b_1^* X)}} \quad (4)$$

 doi:10.20944/preprints202407.1325.v1

Where, p(X) is the predicted output b_0 is the intercept term and b_1 is the coefficient of the single input value x. The logistic function transforms the linear combination of the predictors into a probability value between 0 and 1.

3.1.6. Multi-Layer Perceptron

MLP is a class of feedforward artificial neural networks. It consists of multiple layers of nodes (neurons) connected in a directed graph, where each layer is fully connected to the next one. MLPs are capable of learning complex patterns through supervised learning and are commonly used for tasks such as classification, regression, and function approximation [57].

Input Layer: Let $X = (x_1 + \cdots + x_n)$ be the input vector. Hidden Layer: For the l^{th} hidden layer, let $h_i^{(l)}$ be the activation of i^{th} neuron. The activation is computed as:

$$h_i{}^{(l)} = \emptyset^{(l)}\left(\sum_j \omega_{ij}^{(l)} x_j + b_i^{(l)}\right) \quad (5)$$

Where, $\emptyset^{(l)}$ is the activation function for the l^{th} layer.

$\omega_{ij}^{(l)}$ is the weight connecting the j^{th} neuron in the $(l-1)^{th}$ layer to the i^{th} neuron in the l^{th} layer.

$b_i^{(l)}$ is the bias term for the i^{th} neuron in the l^{th}layer.

$h_i{}^{(l-1)}$ is the activation of the j^{th} neuron in the $(l-1)^{th}$ layer. For the input layer, $h_j{}^{(o)} = x_j$.

Output Layer: Let y^k be the output of the k^{th}neuron in the output layer. It is computed as:

$$y_k = \emptyset^{(L)}\left(\sum_j \omega_{kj}^{(L)} h_j{}^{(L-1)} + b_k^{(L)}\right) \quad (6)$$

We distinguish $\emptyset^{(1)}$ and $\emptyset^{(2)}$ because different layers may have different activation functions.

where:

- L is the index of the output layer.
- $\phi(L)$ is the activation function for the output layer.

3.1.7. AdaBoost

AdaBoost (Adaptive Boosting) is an ensemble learning algorithm that combines the predictions of multiple weak classifiers to create a strong classifier. It works by focusing on the training samples that are hard to classify and adjusting the weights of the weak classifiers accordingly [58].

Given a training set $T = \{T = (x_i, y_i)\}_{i=1}^N$, where x_i are the input feature and y_i are the class labels $(y_i \in \{-1,1\})$ the AdaBoost algorithm can be described as follows:

1. Initialize the Weights:
   $w_i^{(1)} = \frac{1}{N}$ for all i=1,2,…,N
2. For t=1 to T (number of iterations):
   a. Train a Weak Classifier h_t using the weighted training set.
   b. Compute the Weighted Error ϵt :

$$\epsilon_t = \frac{\sum_{i=1}^N w_i^{(t)} * I(h_t(x_i) \neq y_i)}{\sum_{i=1}^N w_i^{(t)}} \quad (7)$$

where I is the indicator function that returns 1 if the condition is true and 0 otherwise.

   c. Compute the Classifier Weight $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$

   d. Update the Weights:

$$w_i^{(t+1)} = w_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right)$$

   Normalize the weights so that they sum to 1:

 doi:10.20944/preprints202407.1325.v1

$$w_i^{(t+1)} = \frac{\omega_i^{(t+1)}}{\sum_{j=1}^{N} w_j^{(t+1)}}$$

3. Final Strong Classifier:

The strong classifier is a weighted majority vote of the T weak classifiers:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t \, h_t(x)) \quad (8)$$

3.1.8. Sentiment Analysis

Sentiment analysis is the process of classifying text into sentiment categories such as positive, negative [16,17]. There are two popular approaches to perform sentiment analysis namely, lexicon-based approach and ML based approach. The former approach requires the use of corpus or dictionary to map words according to their semantic orientation into sentiment categories. Whilst the latter approach involves the use of labeled samples to train the ML algorithm for prediction. In this context, due to the unavailability of labelled sent for training purpose, we consider the use of lexicon-based sentiment analysis suitable. We adopted Afinn developed in [59] based on usage and good performances shown across several sentiment analysis tasks. Table 1 below presents the statistics of Afinn.

Table 1. Afinn statistics.

Label	Count	Example
Positive	878	Good
Negative	1598	Cry

3.2. Datasets

This study adopted three datasets to account for bias of algorithms towards a particular dataset. For identification purpose, we rename the datasets as D1 and D2. Subsequent sections will provide a summary of the datasets.

3.2.1. Dataset 1 (D1)

The dataset D1 was crawled from Reddit [60] and were annotated by two domain experts with labels namely, not depressed, moderate and severe. The label *"not depressed"*, indicate the text shows no sign of depression, while *"moderate"* and *"severe"* indicate this is a depressive text and the strength of depression in text with *"severe"* being high. The dataset has been used in [61–63]. Specifically, was used in the competition of the Second Workshop on Language Technology for Equality, Diversity, and Inclusion [64]. Authors in [64] pre-processed the data by removing the duplicates to create the final dataset (10,251 posts) used in this study. The dataset is publicly available and can be accessed via this link https://github.com/rafalposwiata/depression-detection-lt-edi-2022/tree/main/data (accessed 12th November, 2023). Table 2 below shows the summary of the data.

Table 2. D1 statistics.

Label	$Count_{raw}$	$Count_{preprocessed}$	Example
Not depressed	4,649	3503	Happy New Years Everyone: We made it another year
Moderate	10,494	5780	My life gets worse every year. That's what it feels like anyway
Severe	1,489	968	Words can't describe how bad I feel right now: I just want to fall asleep forever.

 doi:10.20944/preprints202407.1325.v1

3.2.2. Datasets 2 (D2)

The dataset D2, was prepared by the authors in [35]. In their study, they reconstructed the dataset of [65] obtained from Reddit into 4 depression severity levels covering the clinical depression standards on social media. The dataset was labelled manually by two human annotators and an additional annotator was employed for cases with label disagreement. Table 3 below shows the data statistics.

Table 3. D2 statistics.

Label	$Count_{raw}$	Example
Minimal	2587	I just got out of a four year, mostly on but sometimes off relationship. The last interaction we had; he was moving out. The night before, he had strangled me. We've had a toxic relationship, but mostly loving. He truly tried to love me as much as possible but would get drunk and be verbally abusive.
Mild	290	I just feel like the street life has fucked my head up. There's so much I don't even know how to talk about anymore, I just hold that shit. The only person I can really chat with is a pal I know at the bar. He has PTSD and shit from the military bad, hard-up alcoholic nowadays after killing people. We talk once every few weeks and we are open and it's cool. But normal people?
Moderate	394	Sometimes, when I finally got out of bed and stood up, I felt like "Ugh, *finally*". Still, it did not happen every morning, and even when it did, I still felt rested from the long sleep, so I thought no more of it. Also, they were never nightmares. Sadly, my body got habituated to the sleep-component of Mirtazapine after about five months, and my old, warped sleep cycle slowly creeped back into my life. The only benefit left in the medicine was the mild mental cushioning it provided, but at the same time I started to suspect that what I needed wasn't cushioning but to make new constructive life decisions, that only I could make.
Severe	282	I know that I can't be unemployed forever but I'm just too anxious to really do anything. And everyone in my family keeps asking what my plan is and I keep lying because saying I've got nothing is just too humiliating. I'm just stuck. Have any of you have gone through something similar, and have any advice? I appreciate it.

4. Results and Discussion

This section presents the result of our experiments. The evaluation of the algorithms in terms of performance is reported using metrics namely, accuracy (A), precision (P), recall (R), and F1 score (F). Table 4 presents the comparative experimental results of the traditional ML algorithms fitted using the TF-IDF numerical vectors.

 doi:10.20944/preprints202407.1325.v1

Table 4. Evaluation result of ML algorithms.

	D1				D2			
Algorithms	A	P	R	F	A	P	R	F
LR (TF-IDF)	0.37	0.42	0.36	**0.38**	**0.74**	**0.69**	**0.73**	**0.67**
NB (TF-IDF)	0.36	0.40	0.36	0.36	0.72	0.52	0.71	0.60
SVM (TF-IDF)	**0.43**	**0.50**	**0.43**	0.31	0.72	0.56	0.71	0.60
GBM (TF-IDF)	0.39	0.46	0.39	0.35	0.73	0.67	0.72	0.66

In the D1 experimentation, results showed that SVM achieved the best outcome in terms of accuracy, precision, and recall. However, LR showed a better F1 score. It is worth noting that in terms of computational time, SVM took a very long time, GBM took a moderate time, whilst LR and NB was fast. Overall, the traditional ML algorithms performed poorly, most especially with the *"severe"* class. The *"not depressed"* class showed the best F1 score of all the three classes. This is unsurprising considering that the *"severe"* class is the minority class. In the D2 experimentation, results showed that LR outperformed other models. In general, the traditional ML algorithms achieved good results. Across both datasets (D1 & D2), the LR showed an able-to-compete performance. However, the performance was generally poor in D1. Thus, our results showed that the traditional ML models are unable to cater for the diversity and class imbalance in dataset. Furthermore, we performed second experiment, in which we used the word embedding vectors (of BERT & SBERT) and adopted both the traditional ML algorithms (LR, SVM and GBM) and deep learning algorithms (BiLSTM & BiGRU) as classifiers. Results of the second experiment are shown in Table 5 below.

Table 5. Evaluation result of the LLMs.

	D1				D2			
Algorithms	A	P	R	F	A	P	R	F
BERT + LR	0.63	0.63	0.61	0.62	0.72	0.67	0.72	0.69
BERT + SVM	0.65	0.66	0.64	0.63	0.72	0.59	0.72	0.61
BERT + GBM	0.65	0.67	0.65	0.63	0.72	0.64	0.72	0.67
BERT + BiGRU	0.61	0.67	0.61	0.58	0.69	0.68	0.69	0.68
BERT + BiLSTM	0.61	0.68	0.61	0.60	0.69	**0.69**	0.69	0.69
BERT + Ensemble	0.66	0.68	**0.66**	0.64	0.73	0.66	0.73	0.68
SBERT + LR	0.64	0.65	0.64	0.63	0.74	**0.69**	0.74	0.69
SBERT + SVM	0.65	0.66	0.65	0.63	0.74	0.68	0.74	0.66
SBERT + GBM	0.65	0.64	0.63	0.62	0.73	0.65	0.73	0.66

 doi:10.20944/preprints202407.1325.v1

SBERT + BiGRU	0.61	0.63	0.61	0.62	0.71	**0.69**	0.72	**0.70**
SBERT + BiLSTM	0.61	0.62	0.61	0.61	0.73	**0.69**	0.74	**0.70**
SBERT + Ensemble	**0.69**	**0.69**	0.65	**0.68**	**0.76**	**0.69**	**0.75**	**0.70**

In the D1 experimentation, results show that the use of word embedding vectors yield an improvement across all evaluation metrics in D1. However, the models show marginal improvement over each other. Furthermore, our approach performed best in terms of accuracy, precision, and F1 score. In the D2 experimentation, our approach showed improvements in terms of accuracy, recall, and F1 score. However, it is worth stating that improvements in the model performance was marginal when compared to the first experimental results presented in Table 4. This can be attributed to the proportionate distribution of the classes except the *"minimal"* class. Across the datasets (D1 & D2), our approach produced the best performance. This evidences the consistency, suitability, and efficiency of our approach.

Results from Table 6 show that incorporating sentiment analysis as a feature enhances the model performance of the ensemble models. However, the performance of the ML algorithms when used as a stand-alone classifier with the embeddings yielded similar results to the second experimental results presented in Table 5. Overall, our approach performed best in D1 in terms of accuracy, recall, and F1 score. However, showed the best results across all metrics in D2. This further evidence the robustness, consistency, and efficiency of our approach. In comparison to other studies, our approach shows better performance to existing results as shown in Tables 7 and 8 below.

Table 6. Evaluation results using LLMs (with sentiment analysis).

	D1				D2			
Algorithms	A	P	R	F	A	P	R	F
BERT + LR_{AFINN}	0.63	0.64	0.63	0.63	0.66	0.71	0.68	0.70
BERT + SVM_{AFINN}	0.66	**0.72**	0.66	0.62	0.73	0.67	0.72	0.66
BERT + GBM_{AFINN}	0.65	0.67	0.66	0.63	0.72	0.66	0.72	0.67
BERT + $BiGRU_{AFINN}$	0.65	0.68	0.64	0.61	0.69	0.66	0.67	0.68
BERT + $BiLSTM_{AFINN}$	0.65	0.67	0.64	0.65	0.72	0.70	0.73	0.71
BERT + $Ensemble_{AFINN}$	0.71	0.69	0.65	0.67	0.74	0.65	0.71	0.67
SBERT + LR_{AFINN}	0.64	0.65	0.63	0.64	0.75	0.71	**0.74**	0.72
SBERT + SVM_{AFINN}	0.65	0.65	0.65	0.63	0.74	0.72	0.70	0.66
SBERT + GBM_{AFINN}	0.64	0.65	0.64	0.63	0.73	0.65	0.72	0.67
SBERT + $BiGRU_{AFINN}$	0.60	0.61	0.58	0.60	0.71	0.66	0.70	0.68
SBERT + $BiLSTM_{AFINN}$	0.60	0.61	0.58	0.59	0.73	0.70	0.72	0.71
SBERT + $Ensemble_{AFINN}$	**0.74**	0.71	**0.68**	**0.69**	**0.83**	**0.77**	**0.74**	**0.76**

 doi:10.20944/preprints202407.1325.v1

XLNet + Ensemble$_{AFINN}$	0.67	0.70	**0.68**	0.66	0.75	0.67	0.72	0.71
ALBERT + Ensemble$_{AFINN}$	0.67	0.68	0.66	0.64	0.72	0.64	0.71	0.70
RoBERTa + Ensemble$_{AFINN}$	0.69	0.68	0.66	0.67	0.75	0.67	0.72	0.71

Table 7. Performance comparison to previous studies (D1).

	D1			
Literature	A	P	R	F
Poswiata & Perelkiewicz [63]	0.69	0.66	0.62	0.63
Muñoz, & Iglesias [60]	-	0.58	0.61	0.58
Our study	**0.74**	**0.71**	**0.68**	**0.69**

Table 8. Performance comparison to previous studies (D1).

	D2			
Literature	A	P	R	F
Muñoz & Iglesias [60]	-	0.75	0.73	0.74
Ilias et al. [66]	-	0.73	0.73	0.73
Our study	0.83	0.77	0.74	0.76

5. Conclusions

This paper aims to develop a depression detection model using a well performing ML algorithm. To this end, we compared several ML algorithms ranging from the traditional ML to the deep learning algorithms. Furthermore, we utilised TF-IDF and worde embeddings from the LLMs to generate the numeric vectors fitted into the ML models. Specifically, we compared the influence of sentiment indicator as an external feature in the models. Our experimental results evidenced that the use of SBERT and stacking ensemble model achieved SOTA results. Our result is beneficial to the medical/healthcare stakeholders and practitioners for early depression detection. This intervention helps in making a more accurate diagnoses and monitoring the symptoms overtime.

Implications, Limitations and Future Work

Theoretically, we evidence that sentiment analysis indicator is an important feature for enhancing the performance of depression detection. Secondly, we evidence that standalone traditional ML algorithm are limited in generalisation as they often do not perform well across different datasets. We evidence that the words embedding models from the transformer architectures yield an able to compete performance across several datasets. Thus, can be relied on as an off-the-shelf depression detection model.

We have presented the use of stacked ensemble ML models with SBERT and sentiment indicator for the prediction of depression, with improved metrics when compared to previous works. However, SBERT as with any BERT model can struggle to generate embeddings when sentences become too long. Also, it may struggle with context-aware tasks, where there might be subtle semantics difference in sentences. To this end, in future works, we will experiment with newer larger language models such as GPT, LLaMa and PaLM variants with larger trainable parameters as recent research [67] has shown to have better sentiment score when compared to BERT. Also, we will explore the explainability of the black box SBERT model. For example, we need to understand how the Siamese BERT models update their weight during training having trained with different input sentence but uses same weight for each training epoch. Furthermore, it is worth stating that our experimental results relied on Reddit datasets, as such, we envisage evaluating our methodology with datasets from other social networks. Similarly, we recommend extending our methodology in the detection of harmful contents online and the detection of academic stress.

 doi:10.20944/preprints202407.1325.v1

Author Contributions: Conceptualization, B.O.; methodology, B.O., H.S., and O.S.; software, B.O.; validation, B.O. and O.S.; formal analysis, B.O.; investigation, B.O.; resources, B.O.; data curation, B.O.; writing—original draft preparation, B.O., H.S., and O.S.; writing—review and editing, B.O., H.S., and O.S.; visualization, B.O., H.S., and O.S.; supervision, B.O.; project administration, B.O., O.S., funding acquisition, O.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data are presented in the study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Figuerêdo, J.S.L., Maia, A.L.L. and Calumby, R.T., **2022**. Early depression detection in social media based on deep learning and underlying emotions. *Online Social Networks and Media*, *31*, p.100225.
2. Thapar, A., Eyre, O., Patel, V. and Brent, D., **2022**. Depression in young people. *The Lancet*, *400*(10352), pp.617-631.
3. World Health Organization., 2023. Depressive disorder (depression). https://www.who.int/en/news-room/fact-sheets/detail/depression (accessed 27th August, 2023)
4. Cai, Y., Wang, H., Ye, H., Jin, Y. and Gao, W., **2023**. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, *217*, p.119538.
5. World health Organization, **2017**. Depression and other common mental disorders: global health estimates. Technical Report. World Health Organization. (Accessed 19th September 2023) https://apps.who.int/iris/handle/10665/254610.
6. Zhang, T., Yang, K., Alhuzali, H., Liu, B. and Ananiadou, S., **2023**. PHQ-aware depressive symptoms identification with similarity contrastive learning on social media. *Information Processing & Management*, *60*(5), p.103417.
7. Liang, Y., Liu, L., Ji, Y., Huangfu, L. and Zeng, D.D., **2023**. Identifying emotional causes of mental disorders from social media for effective intervention. *Information Processing & Management*, *60*(4), p.103407.
8. Beck, A.T., Ward, C.H., Mendelson, M., Mock, J. and Erbaugh, J., **1961**. An inventory for measuring depression. *Archives of general psychiatry*, *4*(6), pp.561-571.
9. Radloff, L.S., **1991**. The use of the Center for Epidemiologic Studies Depression Scale in adolescents and young adults. *Journal of youth and adolescence*, *20*(2), pp.149-166.
10. Kovacs, M., **1992**. Children's depression inventory. *Toronto Ontario*.
11. Angold, A. and Costello, E.J., **1987**. Mood and feelings questionnaire (MFQ). *Durham: Developmental Epidemiology Program, Duke University*. https://devepi.duhs.duke.edu/measures/the-mood-andfeelings-questionnaire-mfq/ (accessed October 28, 2023).
12. Kroenke, K., Spitzer, R.L. and Williams, J.B., **2001**. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, *16*(9), pp.606-613.
13. Chorpita, B.F., Moffitt, C.E. and Gray, J., **2005**. Psychometric properties of the Revised Child Anxiety and Depression Scale in a clinical sample. *Behaviour research and therapy*, *43*(3), pp.309-322.
14. Epstein, R.M., Duberstein, P.R., Feldman, M.D., Rochlen, A.B., Bell, R.A., Kravitz, R.L., Cipri, C., Becker, J.D., Bamonti, P.M. and Paterniti, D.A., **2010**. "I didn't know what was wrong:" how people with undiagnosed depression recognize, name and explain their distress. *Journal of general internal medicine*, *25*, pp.954-961.
15. Boerema, A.M., Kleiboer, A., Beekman, A.T., van Zoonen, K., Dijkshoorn, H. and Cuijpers, P., **2016**. Determinants of help-seeking behavior in depression: a cross-sectional study. *BMC psychiatry*, *16*, pp.1-9.
16. Ogunleye, B.O., **2021**. *Statistical learning approaches to sentiment analysis in the Nigerian banking context*. Sheffield Hallam University (United Kingdom).
17. Ogunleye, B., Brunsdon, T., Maswera, T., Hirsch, L. and Gaudoin, J., **2023**, August. Using Opinionated-Objective Terms to Improve Lexicon-Based Sentiment Analysis. In *International conference on soft computing*

for problem-solving (pp. 1-23). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-97-3292-0_1

18. Chancellor, S. and De Choudhury, M., **2020**. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, *3*(1), p.43.
19. Pérez, A., Parapar, J., Barreiro, Á. and Lopez-Larrosa, S., **2023**, July. Bdi-sen: A sentence dataset for clinical symptoms of depression. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2996-3006).
20. Wang, Y., Wang, Z., Li, C., Zhang, Y. and Wang, H., **2022**. Online social network individual depression detection using a multitask heterogenous modality fusion approach. *Information Sciences*, *609*, pp.727-749.
21. Islam, M.R., Kamal, A.R.M., Sultana, N., Islam, R. and Moni, M.A., **2018**, February. Detecting depression using k-nearest neighbors (knn) classification technique. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.
22. Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S. and Goharian, N., **2018**. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.
23. Bierbaum, J., Lynn, M. and Yu, L., **2022**, April. Utilizing Pattern Mining and Classification Algorithms to Identify Risk for Anxiety and Depression in the LGBTQ+ Community During the COVID-19 Pandemic. In *Companion Proceedings of the Web Conference 2022* (pp. 663-672).
24. Skaik, R. and Inkpen, D., **2020**, December. Using twitter social media for depression detection in the canadian population. In *Proceedings of the 2020 3rd Artificial Intelligence and Cloud Computing Conference* (pp. 109-114).
25. Hosseini-Saravani, S.H., Besharati, S., Calvo, H. and Gelbukh, A., **2020**, October. Depression detection in social media using a psychoanalytical technique for feature extraction and a cognitive based classifier. In *Mexican International Conference on Artificial Intelligence* (pp. 282-292). Cham: Springer International Publishing.
26. He, L., Chan, J.C.W. and Wang, Z., **2021**. Automatic depression recognition using CNN with attention mechanism from videos. *Neurocomputing*, *422*, pp.165-175.
27. Ive, J., Gkotsis, G., Dutta, R., Stewart, R. and Velupillai, S., **2018**, June. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic* (pp. 69-77).
28. Amanat, A., Rizwan, M., Javed, A.R., Abdelhaq, M., Alsaqour, R., Pandya, S. and Uddin, M., **2022**. Deep learning for depression detection from textual data. *Electronics*, *11*(5), p.676.
29. Almars, A.M., **2022**. Attention-Based Bi-LSTM Model for Arabic Depression Classification. *Computers, Materials & Continua*, *71*(2).
30. Liu, T., Jain, D., Rapole, S.R., Curtis, B., Eichstaedt, J.C., Ungar, L.H. and Guntuku, S.C., **2023**, April. Detecting symptoms of depression on reddit. In *Proceedings of the 15th ACM Web Science Conference 2023* (pp. 174-183).
31. Harrigian, K., Aguirre, C. and Dredze, M., **2020**, November. Do models of mental health based on social media data generalize?. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 3774-3788).
32. Ogunleye, B. and Dharmaraj, B., **2023**. The use of a large language model for cyberbullying detection. *Analytics*, *2*(3), pp.694-707.
33. Cheng, Q., Li, T.M., Kwok, C.L., Zhu, T. and Yip, P.S., **2017**. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. *Journal of medical internet research*, *19*(7), p.e243.
34. Shrestha, A., Tlachac, M.L., Flores, R. and Rundensteiner, E.A., **2022**, September. Bert variants for depression screening with typed and transcribed responses. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers* (pp. 211-215).
35. Naseem, U., Dunn, A.G., Kim, J. and Khushi, M., **2022**, April. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022* (pp. 2563-2572).
36. Monreale, A., Iavarone, B., Rossetto, E. and Beretta, A., **2022**, April. Detecting addiction, anxiety, and depression by users psychometric profiles. In *Companion Proceedings of the Web Conference 2022* (pp. 1189-1197).

37. Sen, I., Quercia, D., Constantinides, M., Montecchi, M., Capra, L., Scepanovic, S. and Bianchi, R., **2022**. Depression at work: exploring depression in major US companies from online reviews. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), pp.1-21.
38. Wu, J., Wu, X., Hua, Y., Lin, S., Zheng, Y. and Yang, J., **2023**, April. Exploring social media for early detection of depression in covid-19 patients. In *Proceedings of the ACM Web Conference 2023* (pp. 3968-3977).
39. Villatoro-Tello, E., Ramírez-de-la-Rosa, G., Gática-Pérez, D., Magimai.-Doss, M. and Jiménez-Salazar, H., **2021**, October. Approximating the mental lexicon from clinical interviews as a support tool for depression detection. In *Proceedings of the 2021 international conference on multimodal interaction* (pp. 557-566).
40. Liu, Y., Kang, K.D. and Doe, M.J., **2022**. Hadd: High-accuracy detection of depressed mood. *Technologies*, *10*(6), p.123.
41. Malik, A., Shabaz, M. and Asenso, E., **2023**. Machine learning based model for detecting depression during Covid-19 crisis. *Scientific African*, *20*, p.e01716.
42. Gallegos Salazar, L.M., Loyola-Gonzalez, O. and Medina-Perez, M.A., **2021**. An explainable approach based on emotion and sentiment features for detecting people with mental disorders on social networks. *Applied Sciences*, *11*(22), p.10932.
43. Burdisso, S.G., Errecalde, M. and Montes-y-Gómez, M., **2019**. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, *133*, pp.182-197.
44. Trotzek, M., Koitka, S. and Friedrich, C.M., **2018**. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, *32*(3), pp.588-601.
45. Adarsh, V., Kumar, P.A., Lavanya, V. and Gangadharan, G.R., **2023**. Fair and explainable depression detection in social media. *Information Processing & Management*, *60*(1), p.103168.
46. Guo, Z., Ding, N., Zhai, M., Zhang, Z. and Li, Z., **2023**. Leveraging domain knowledge to improve depression detection on Chinese social media. *IEEE Transactions on Computational Social Systems*, *10*(4), pp.1528-1536.
47. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., **2018**. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
48. Reimers, N. and Gurevych, I., **2019**. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
49. Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M. and Torres, J.F., **2018**. Stacking ensemble learning for short-term electricity consumption forecasting. *Energies*, *11*(4), p.949.
50. Kwon, H., Park, J. and Lee, Y., **2019**. Stacking ensemble technique for classifying breast cancer. *Healthcare informatics research*, *25*(4), pp.283-288.
51. Rajagopal, S., Kundapur, P.P. and Hareesha, K.S., **2020**. A stacking ensemble for network intrusion detection using heterogeneous datasets. *Security and Communication Networks*, *2020*(1), p.4586875.
52. Charoenkwan, P., Chiangjong, W., Nantasenamat, C., Hasan, M.M., Manavalan, B. and Shoombuatong, W., **2021**. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Briefings in bioinformatics*, *22*(6), p.bbab172.
53. Akyol, K., **2020**. Stacking ensemble based deep neural networks modeling for effective epileptic seizure detection. *Expert Systems with Applications*, *148*, p.113239.
54. Ribeiro, M.H.D.M. and dos Santos Coelho, L., **2020**. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied soft computing*, *86*, p.105837.
55. Natekin, A. and Knoll, A., **2013**. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, p.21.
56. Saini, D., Chand, T., Chouhan, D.K. and Prakash, M., **2021**. A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on X-ray images. *Biocybernetics and Biomedical Engineering*, *41*(2), pp.419-444.
57. Grosse, R., **2019**. Lecture 5: Multilayer Perceptrons. *inf. téc.*
58. Tsai, J.K. and Hung, C.H., **2021**. Improving AdaBoost classifier to predict enterprise performance after COVID-19. *Mathematics*, *9*(18), p.2215.
59. Nielsen, F.Å., 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

60. Sampath, K. and Durairaj, T., **2022**, March. Data set creation and empirical analysis for detecting signs of depression from social media postings. In *International Conference on Computational Intelligence in Data Science* (pp. 136-151). Cham: Springer International Publishing.
61. Muñoz, S. and Iglesias, C.Á., **2023**. Detection of the Severity Level of Depression Signs in Text Combining a Feature-Based Framework with Distributional Representations. *Applied Sciences*, *13*(21), p.11695.
62. Shi, Y., Tian, Y., Tong, C., Zhu, C., Li, Q., Zhang, M., Zhao, W., Liao, Y. and Zhou, P., **2023**, November. Detect depression from social networks with sentiment knowledge sharing. In *Chinese national conference on social media processing* (pp. 133-146). Singapore: Springer Nature Singapore.
63. Tavchioski, I., Robnik-Šikonja, M. and Pollak, S., **2023**. Detection of depression on social networks using transformers and ensembles. *arXiv preprint arXiv:2305.05325*.
64. Poświata, R. and Perełkiewicz, M., **2022**, May. OPI@ LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 276-282).
65. Turcan, E. and McKeown, K., **2019**. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133*.
66. Ilias, L., Mouzakitis, S. and Askounis, D., **2023**. Calibration of transformer-based models for identifying stress and depression in social media. *IEEE Transactions on Computational Social Systems*, *11*(2), pp.1979-1990.
67. Shobayo, O., Sasikumar, S., Makkar, S. and Okoyeigbo, O., **2024**. Customer Sentiments in Product Reviews: A Comparative Study with GooglePaLM. *Analytics*, *3*(2), pp.241-254.