A textual-based featuring approach for depression detection using machine learning classifiers and social media texts

Raymond Chiong, Gregorius Satia Budhi, Sandeep Dhakal, Fabian Chiong

PII: S0010-4825(21)00293-6

DOI: https://doi.org/10.1016/j.compbiomed.2021.104499

Reference: CBM 104499

To appear in: *Computers in Biology and Medicine*

Received Date: 12 February 2021

Revised Date: 12 May 2021

Accepted Date: 14 May 2021

A textual-based featuring approach for depression detection using machine learning classifiers and social media texts

Raymond Chiong[a,*], Gregorius Satia Budhi[a,b,*] , Sandeep Dhakal[a] and Fabian Chiong[c]
[a] School of Electrical Engineering and Computing, The University of Newcastle, Callaghan, NSW 2308, Australia
[b] Informatics Department, Petra Christian University, Surabaya 60236, Indonesia
[c] Alice Springs Hospital, The Gap, NT 0870, Australia

Abstract
Depression is one of the leading causes of suicide worldwide. However, a large percentage of cases of depression go undiagnosed and, thus, untreated. Previous studies have found that messages posted by individuals with major depressive disorder on social media platforms can be analysed to predict if they are suffering, or likely to suffer, from depression. This study aims to determine whether machine learning could be effectively used to detect signs of depression in social media users by analysing their social media posts—especially when those messages do not explicitly contain specific keywords such as 'depression' or 'diagnosis'. To this end, we investigate several text preprocessing and textual-based featuring methods along with machine learning classifiers, including single and ensemble models, to propose a generalised approach for depression detection using social media texts. We first use two public, labelled Twitter datasets to train and test the machine learning models, and then another three non-Twitter depression-class only datasets (sourced from Facebook, Reddit, and an electronic diary) to test the performance of our trained models in other social media sources. Experimental results indicate that the proposed approach is able to effectively detect depression via social media texts even when the training datasets do not contain specific keywords (such as 'depression' and 'diagnose'), as well as when unrelated datasets are used for testing.

Keywords: *depression detection, social media, textual-based featuring, machine learning, imbalanced data*

INTRODUCTION

According to the World Health Organisation (WHO), depression is the most prevalent mental disorder that affects more than 300 million people worldwide [1]. Depression is also the leading cause of more than two-thirds of suicides every year [2]. However, due to self-denial among some patients and poor recognition of the issue in many places, depression can remain undiagnosed or untreated. The lack of diagnosis and treatment can further aggravate the condition [3], which could lead to reduced quality of life and, in acute cases, an inability to maintain employment [4,5].

Numerous studies in the literature agree that social media platforms, where people freely share their thoughts and express their feelings, could be a vital source for monitoring health issues and trends [6,7]. Posts on platforms, such as Twitter and Facebook, enable researchers to investigate multiple aspects of

[*] Corresponding author. Tel.: (02) 4921 7367
Email addresses: raymond.chiong@newcastle.edu.au (Raymond Chiong), gregorius.satiabudhi@uon.edu.au; greg@petra.ac.id (Gregorius Satia Budhi), sandeep.dhakal@newcastle.edu.au (Sandeep Dhakal), fabian.chiong@nt.gov.au (Fabian Chiong)

psychological concerns and human behaviour [8,9]. Studies focusing on mental depression, for example, have found that tweets posted by individuals with a major depressive disorder could be utilised to predict future episodes of depression in those individuals [3]. A recent survey indicated that an increasing number of people with depression symptoms, especially teenagers and young adults, are turning to social media to express their feelings (see Fig. 1) [10]. Related work in this domain, however, often relies on specific keywords like 'depression' and 'diagnose' when utilising the data, while the fact is that social media users suffering from depression are not likely to use such words directly.

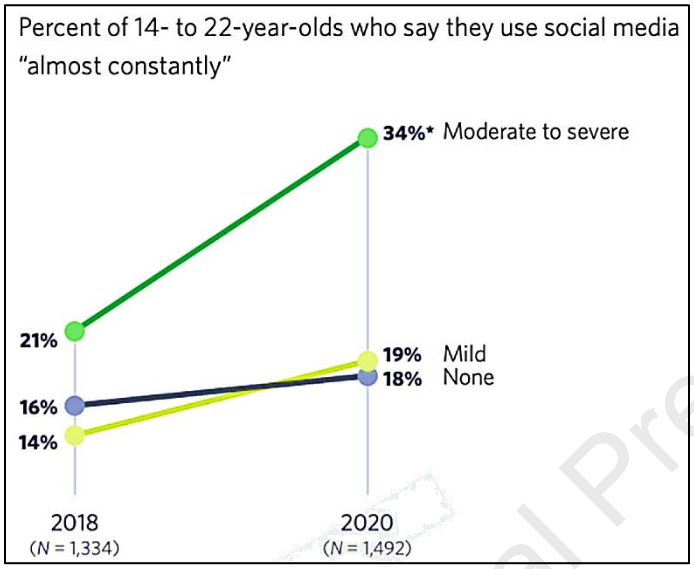


Fig. 1 Frequency of social media use, by depressive symptom levels, 2018 – 2020 [10]

Given the prevalence of depression, its impacts and the potential of using social media texts for predicting depression, the main goal of this study is to analyse whether machine learning (ML) methods could be effectively used to detect depression in people by analysing their social media texts, but without relying on specific keywords. Social media texts are often unstructured, and therefore ML, which is good at dealing with nonlinearity, is expected to be a better option than traditional statistical methods for analysing them. Even though we train the ML models using Twitter posts in this study, they should be equally applicable to other text-based messages. Specifically, the models are first trained and tested using two publicly labelled datasets comprising of Twitter tweets [11,12]—with and without specific keywords such as 'depression' and 'diagnose'. The models are then re-tested using additional non-Twitter depression-class only datasets [13-15], primarily to investigate the performance of our trained models in detecting depression in people who posted messages on various social media platforms. More precisely, the performance of our generalised approach is investigated against five different, publicly available, datasets comprising of social-media text from various sources, including Twitter, Facebook, Reddit, and a personal electronic diary. This approach contrasts with most approaches in the literature that use parts of the same datasets for both training and testing the ML models, as well as relying on specific keywords.

Our detection model is constructed using ML methods combined with textual-based featuring; textual-based featuring extracts the input features from the text itself and is, thus, more independent of the system than other featuring processes [16]. We apply ML classifiers—both single and ensemble models—that are widely used in solving prediction problems; these methods have been selected because of their excellent prediction performance in previous studies [16,17]. The performance of these models is, however, dependent on the type of data and features used for training them. Therefore, we investigate the performance of our models using datasets from several social media texts to verify that the proposed approach is a general one. Additionally, the features used for training are preprocessed and extracted using a combination of various methods that performed well in previous work [16-18]. More specifically, text preprocessing methods, such as tokenisation, stop word removal, detection of negation words, correction of elongation words, and part of speech (POS) lemmatisation, are applied. Regarding the input features, we implement a bag-of-words (BOW) feature extraction method combined with count vectorisation and n-gram words (from unigram to trigram) to extract features. This work thus contributes to the relevant research areas of ML and natural language processing as well as the study of mental health problems by proposing a generalised approach for depression detection that is effective especially when the social media users are not aware of their depression or are in denial.

The rest of this paper is organised as follows. In Section 2, we review related work on depression detection. Then in Section 3, we describe, in detail, the design of the textual-based depression detection framework developed for this study. Experimental results and discussions are presented in Section 4, and Section 5 concludes the paper and highlights potential future research directions.

RELATED WORK

Numerous studies on automatic detection of the symptoms of depression have been carried out using artificial intelligence methods such as ML. A major stream of research involves depression detection using medical data, such as fMRI signature [19], results of depression questionnaires (such as DASS21 and DASS42) [1,20], or clinical criteria for depression as defined in DSM-5 and ICD-10 [21]. Data from clinical interviews, using systems such as Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ [22]) [2,23], has also been collected. Data from DAIC-WOZ includes videos, speeches, and text transcriptions of the participants, who could be either distressed or non-distressed. In recent years, researchers have also focused on depression detection using text messages from social media platforms, such as Twitter, Facebook, Reddit, and WeChat [3,6,9,11,24-27]—in the hope that social media texts can help detect depression even when the individual is unaware of their depression or is in denial.

The majority of research studies on depression detection using social media messages usually follow either a textual-based featuring approach or a person descriptive-based featuring approach. Textual-based featuring focuses on the linguistic features of the social media text, such as words, POS, n-gram, and other linguistic characteristics [3,6,9,25-28]. In contrast, the descriptive-based featuring approach focuses on descriptions of the subject, such as age, gender, employment status, income, consumption of drugs or alcohol, smoking, and other details of the subject or patient [11,21,29-32]. These features are then input into the detection models. Most models for depression detection have been developed using ML classifiers, such as the Support Vector Machine (SVM), Multilayer Perceptron (MLP), Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), Maximum Entropy (ME), K-Nearest Neighbours (KNN), Adaptive Boosting (AB), Random Forest (RF), Gradient Boosting (GB), Bagging Predictors (BP), and other single and ensemble models [1,3,6,19,21,26,29,30]. Deep learning methods, such as the Long Short-Term Memory (LSTM) [24] and Convolutional Neural Network (CNN) [9,33], have also been used. Additionally, several studies have constructed custom detectors [11,26,31]. A concise overview of related work from the past five years (2017-2021) is provided in Table *1*.

In this paper, we discuss our efforts towards building a depression detection framework using textual features from social media posts. This is achieved by utilising some proven methods of preprocessing, featuring and ML classifiers from our previous studies [16,17]. In research on depression detection using social media texts, custom datasets are frequently created but not made publicly available. In contrast, in this study, we utilise five publicly available datasets. These include two binary-class Twitter datasets that are used for both training and testing, and three single-class datasets from Facebook, Reddit and an electronic diary for further testing. While our goal is a generalised approach for depression detection in social media texts, the datasets include only text messages and exclude any emoticons, emojis, pictures, videos and web links that are commonly part of social media messages. Additionally, we address the issue of overfitting that generally arises when collecting depression data from social media messages. A model might perform poorly on datasets it was not trained on due to overfitting. We also focus our efforts on overcoming the problem of imbalanced data samples, which can negatively impact the performance of classifier models.

Table 1 An overview of related work from 2017 to 2021

Authors	**Year**	**Feature Type**	**Dataset sources**	**Methods**[1]	**Results**[2]
Depression detection on social media					
Shen et al. [11]	2017	Descriptive	Twitter (Public/Pu, [11])	**MDL**[3], MSNL[3], WDL[3], NB	Acc: 85%; Pre: 85%; Rec: 85%; F1: 85%
Hassan et al. [6]	2017	Textual	Twitter (Private/Pr)	**SVM,** NB, ME	Acc: 91%; Pre: 83%; Rec: 79%
Chen et al. [24]	2018	Textual	Survey and WeChat (Pr)	LSTM	Present the results in several graphs
Islam et Al. [25]	2018	Textual	Facebook (Pr)	**DT,** KNN, SVM, Ensemble	Pre: 59%; Rec: 97%; F1: 73%
Burdisso et al. [26]	2019	Textual	Reddit (Pr)	**SS3**[3], KNN, LR, SVM, NB	Pre: 63%; Rec: 60%; F1: 61%
Fatima et al. [27]	2019	Textual	Reddit (Pr)	**MLP,** SVM, LR	Acc: 91.63%; Pre: 91.83%; Rec: 91.85%
Lin et al. [33]	2020	Visual and Textual	Twitter (Pu, [11]) and Images	**CNN**	Acc: 88.4%; Pre: 90.3%; Rec: 87%; F1: 93.6%
Alsagri and Mourad [3]	2020	Textual	Twitter (Pr)	**SVM,** NB, DT	Acc: 82.5%; Pre: 73.91%; Rec: 85%; F1: 79.06%; AUC: 0.78
Kim et al. [9]	2020	Textual	Reddit (Pu, [9])	**CNN,** XGBoost	Acc: 75.13%; Pre: 89.1%; Rec: 71.75%; F1: 79.49%
Depression detection on other sources (non-social media)					
Jung et al. [31]	2017	Textual/ Ontology	35 FAQs about depression from multiple sources	**DT,** LR	Acc: 75% ; Pre: 76.1%
Samareh et al. [23]	2018	Audio, video and textual	DAIC-WOZ[3] (Pu, [22])	**RF**	RMSE: 5.12; MAE: 4.12
Priya et al. [1]	2020	Descriptive	DASS-21[3] (Pr)	**NB,** DT, RF, SVM, KNN	Acc: 85.5%; Pre: 82.2%; Rec: 85%; F1: 83.6
Kumar et al. [20]	2020	Descriptive	DASS-42 (Pr)	**RBFN**[3], NB, KNN, MLP, RF, K-Star, J48	Acc: 96%; Pre: 96%; Rec: 96%; F1: 96%; AUC: 0.99
Srimadhur and Lalitha [2]	2020	Spectrogram & End-to-end	DAIC-WOZ (Pu, [22])	**CNN** (Non-depressed class)	Pre: 65%; Rec: 92%; F1: 76%
Jothi et al. [30]	2020	Descriptive	Online survey (Pr)	**J48,** NB, RF	Acc: 95.7%; Rec: 97.5%; Spe: 86.3%
Filho et al. [29]	2021	Descriptive	Patients' clinical evaluation (Pr)	**RF,** LR, KNN, DT, AB, SVM, GB	Acc: 89%

[1] The first method in bold is the method with the best result

[2] Acc = Accuracy; Pre = Precision; Rec = Recall; Spe = Specificity; F1 = F-measure; AUC = Area Under the Curve

[3] SS3 = Sequential S3 (Smoothness, Significance, and Sanction); MDL = Multimodal Depressive Dictionary Learning; WDL = Wasserstein Dictionary Learning; MSNL = Multiple Social Networking Learning; fMRI = functional Magnetic Resonance Imaging; MLDA = Maximum Entropy Linear Discriminant Analysis; DASS = Depression, Anxiety and Stress Scale questionnaire; RBFN =

Radial Basis Function Network; DSM-5 = Diagnostic and Statistical Manual of Mental Disorders 5th edition; ICD-10 = International Statistical Classification of Diseases and Related Health Problems nu. 10; BCC = Bayesian Classifier Chains; PCC = Probabilistic Classifier Chains; SCC = Super Class Classifier, Bagging, ECC = Ensemble of Classifier Chains; Pruned sets, CC = Classifier Chains; DAIC-WOZ = The Distress Analysis Interview Corpus Wizard of Oz.

METHODS

We designed a framework for detecting depression through an analysis of only social media texts, by using methods inspired by previous studies in the literature. For the preprocessing of the text reviews (see Fig. 2), we implemented several methods that performed well in previous studies [16], as follows:

a. Removal of punctuation, numbers, and stopwords.
b. Word correction, which is an essential step in preprocessing, for reducing the diversity of features. The following three methods of word correction are utilised in this study:
   - *Spelling error correction.* Detectors differentiate between misspelt words and their correct forms, thereby unnecessarily increasing the detection complexity. To avoid this, we detected and corrected misspelt words by utilising Peter Norvig's code for spelling correction [34] based on probability theory. In this method, the investigated words are compared against a large database of words and the most probable replacements are chosen for misspelt words.
   - *Elongation words correction.* Word diversity, and thereby the detection complexity, are also impacted by the presence of elongated words or word-stretchers such as 'yesss', 'fiiine' and 'yoouu'. These were also returned to their original forms using Peter Norvig's code for spelling correction [34].
   - *Negative words correction:* Despite the numerous forms of negative words, their common goal is to introduce negation in the sentence. They were reduced to their basic negative form 'not'.
c. POS tagging and lemmatisation. POS tagging by assigning words to their syntactic functions—such as noun, pronoun, adjective, verb and adverb—puts the words in context [35]. Doing so allows lemmatising the word to the correct context, i.e., changing the word back to its basic form—an essential step in reducing word diversity and making recognition easier.

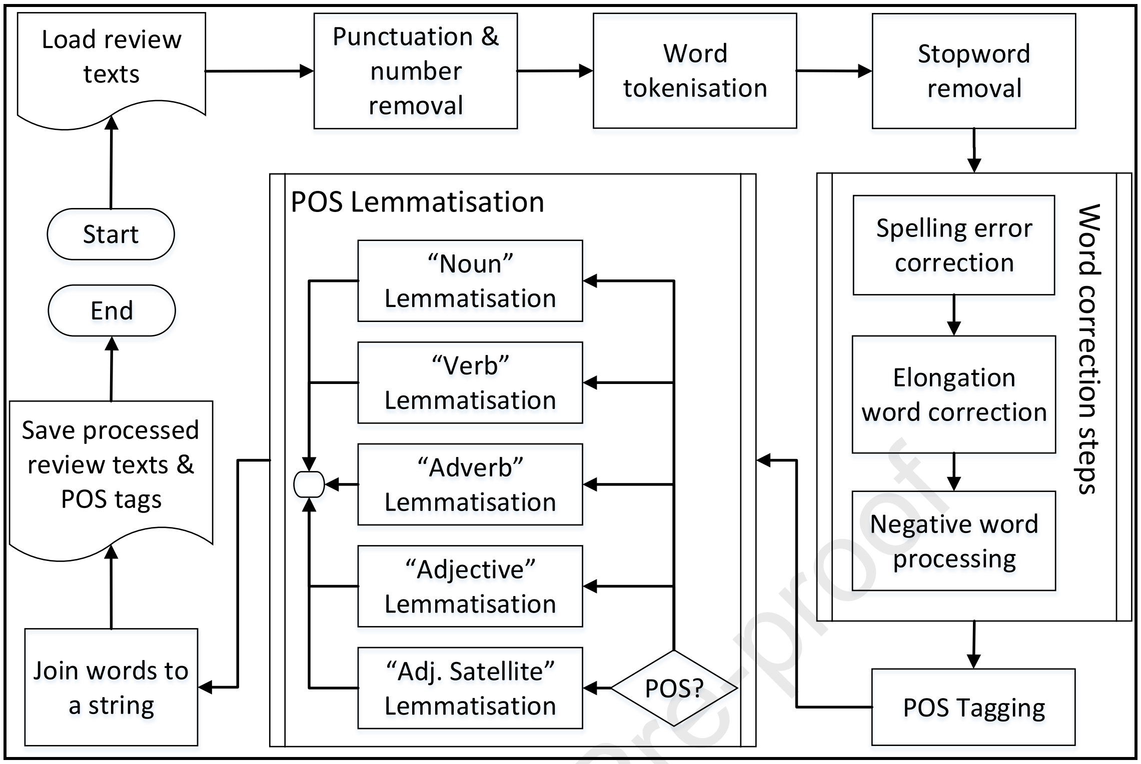
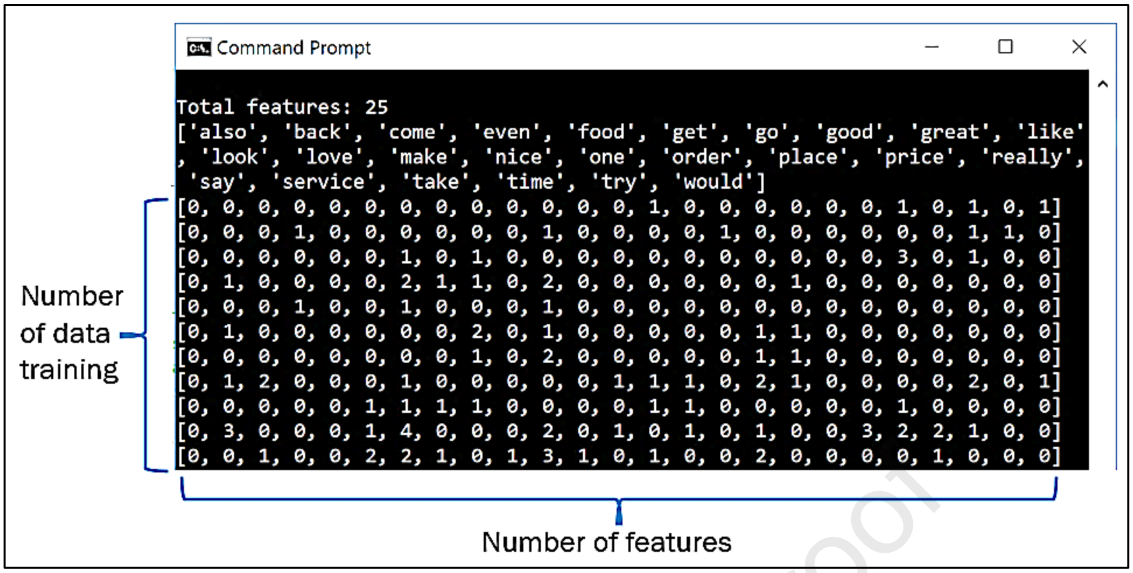


Fig. 2 Preprocessing steps [16]

Since the features are extracted directly from social media messages, we used the BOW feature-extraction method. BOW is commonly used for textual-based features [36] and works by decomposing the entire text into a group of singular words. In this study, we used it in combination with the n-gram technique to capture both singular words and word combinations that convey one meaning but are composed of multiple words. Regarding word-combinations, BOW checks the existence of a contiguous sequence of n words from the given text sample. In this study, we limited the checking to trigrams, since sequences of more than three words are infrequent in real-world texts. After decomposing the text as discussed above, the terms were sorted based on their frequency across the given dataset. A sample screenshot showing how the features are extracted using BOW is shown in Fig. 3. Whereas the screenshot is for an example scenario where the length of features is only 25 unigram words, 5000 unigram to trigram phrases were used for the length of features in our experiments. These settings were determined based on previous studies related to textual-based featuring [16,17].



```
Total features: 25
['also', 'back', 'come', 'even', 'food', 'get', 'go', 'good', 'great', 'like'
, 'look', 'love', 'make', 'nice', 'one', 'order', 'place', 'price', 'really',
 'say', 'service', 'take', 'time', 'try', 'would']
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1]
[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0]
[0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 1, 0, 0]
[0, 1, 0, 0, 0, 0, 2, 1, 1, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
[0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[0, 1, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0]
[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0]
[0, 1, 2, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 2, 1, 0, 0, 0, 0, 2, 0, 1]
[0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[0, 3, 0, 0, 0, 1, 4, 0, 0, 0, 2, 0, 1, 0, 1, 0, 1, 0, 0, 3, 2, 2, 1, 0, 0]
[0, 0, 1, 0, 0, 2, 2, 1, 0, 1, 3, 1, 0, 1, 0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0]
```

Fig. 3 A screenshot showing the features extracted using BoW for a simple hypothetical scenario with 25 unigram words used for the feature length.

One of the datasets used in our study is imbalanced, and this could affect prediction [37]. Hence, we proposed a dynamic sampling process designed to optionally implement dynamic random over or under-sampling to overcome the issue of imbalanced data.

The next step in the framework is to train the ML classifiers for the detection of mental depression. Towards this end, we implemented and tested several single and ensemble classifiers that provided excellent text recognition and detection performance in previous studies [16-18]. Those classifiers are:

1. *Logistic Regression:* A generalised linear model, it was originally developed by Nelder and Wedderburn [38] and further improved by Hastie and Tibshirani [39]. Generalised linear models use non-normal dependent variables and thereby overcome a major limitation of linear models, i.e., continuous and normally distributed dependent variables [40,41]. The dependent variables in LR are either unordered or ordered polytomous, and the independent predictor variables are either interval/ratio or dummy variables [42].
2. *Support Vector Machine:* It is a supervised learning model that classifies new data after learning from training data [43]. It works by separating different classes with a hyperplane and then attempts to maximise the separation distance from the hyperplane. The larger the distance, the lower the error generated by the classifier [44]. In this study, the SVM is used together with the linear kernel (LSVM), which is generally recommended for text classification [45].
3. *Multilayer Perceptron:* A feedforward artificial neural network that uses supervised learning [46,47], the MLP continually computes and updates all the weights in its network to minimise error. In its first phase, called the feed-forward phase, the training data is forwarded to the output layer, following which the difference between this output and the desired target (the error) is backpropagated to update the weights of the network in the second phase [48]. In this study, we utilised the Adam optimiser [49] to increase the performance of the MLP.
4. *Decision Tree:* The DT was developed by Quinlan [50] based on Hunt's algorithm [51], and is a useful tool for exploring the cause-and-effect chain. It builds a tree-like decision model for classification and prediction purposes and is typically used as a base classifier for ensemble models (e.g., BP, RF, and AB).

5. *Random Forest:* The RF is an ensemble of DT predictors where all decision trees are trained independently using random vectors. The strength of trees and their correlation determines the error generalisation of the RF. It is relatively robust to outliers and noise [52].
6. *Adaptive Boosting:* This ensemble combines many weak classifiers iteratively over several rounds *[53]*. Starting with equal weighting for all training data, weights of misclassified training data points are boosted, and a new classifier is created with the new unequal weightings. This procedure is then repeated for a set of classifiers [54].
7. *Bagging Predictors:* This ensemble builds a cluster of several single predictors, which are trained in a bootstrapping process that replicates the training set. Classification is performed using plurality voting [55]. The default BP implementation from scikit-learn, which is used in this study, uses the DT as its base predictor.
8. *Gradient Boosting:* The GB, which comprises gradient boosted regression trees, provides a robust, competitive and interpretable algorithm for classification and regression. In the GB, only a single regression tree is used for binary classification [56].

A flowchart for the sampling process and overall design can be seen in Fig. 4 and Fig. 5, respectively.

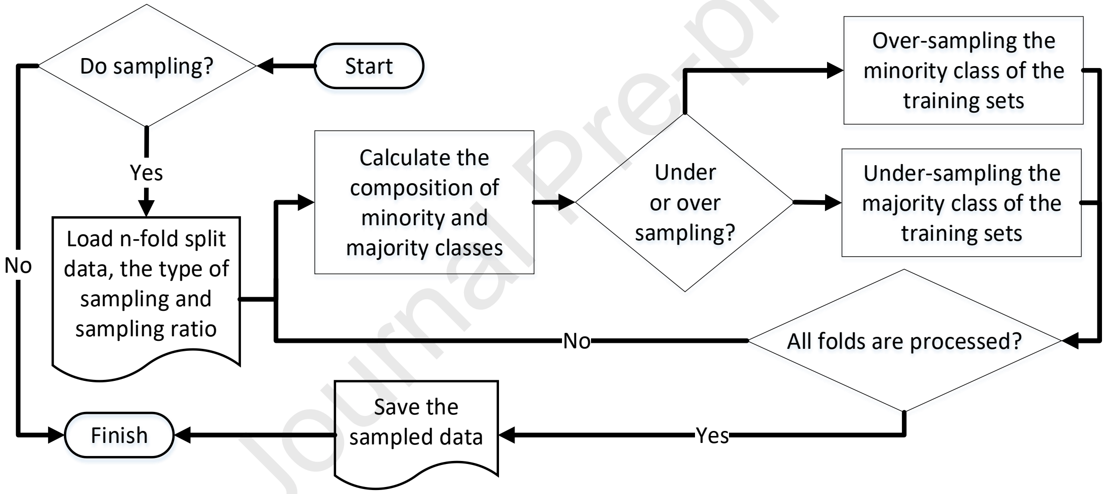

Start
Do sampling?
Yes
No
Load n-fold split data, the type of sampling and sampling ratio
Calculate the composition of minority and majority classes
Under or over sampling?
Over-sampling the minority class of the training sets
Under-sampling the majority class of the training sets
All folds are processed?
No
Yes
Save the sampled data
Finish

Fig. 4 The sampling process

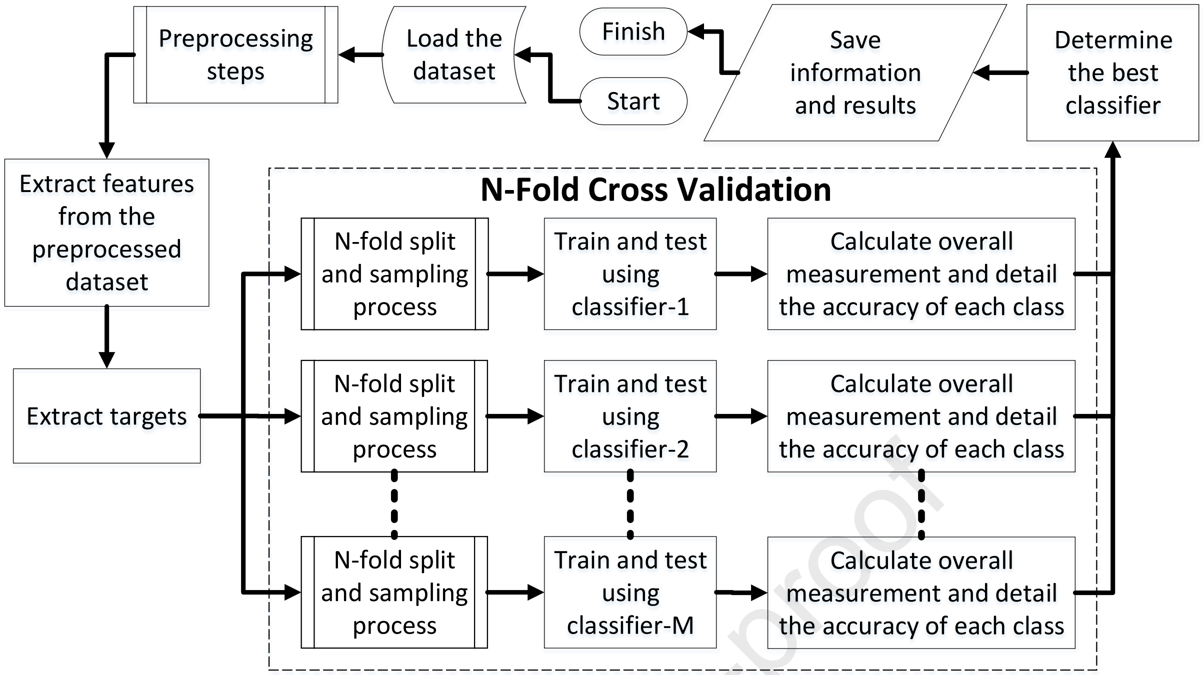

Preprocessing steps
Load the dataset
Finish
Start
Save information and results
Determine the best classifier
Extract features from the preprocessed dataset
Extract targets
N-Fold Cross Validation
N-fold split and sampling process
Train and test using classifier-1
Calculate overall measurement and detail the accuracy of each class
N-fold split and sampling process
Train and test using classifier-2
Calculate overall measurement and detail the accuracy of each class
N-fold split and sampling process
Train and test using classifier-M
Calculate overall measurement and detail the accuracy of each class

Fig. 5 Overall design

We can see from Fig. 5 that all experiments were conducted using the N-fold Cross Validation (CV) method. CV is primarily used in classification and regression models to reduce the bias between an entire dataset and the training/testing sets [57], and it also helps avoid overfitting [58]. In CV, data is split into n disjoint folds or partitions, where n-1 folds (subsamples) are used for training and the remaining fold is used for testing. In this study, we set n = 10; i.e., all experiments were conducted using 10-fold CV. The overall measurements were calculated by averaging the results of each process [59,60].

Four well-known measurements accuracy, precision, recall and F1 (or F-measure) were used to evaluate the performance of the classifiers in terms of depression detection. Measurement components from scikit-learn [61] were implemented, and the relevant equations and functions are provided in Table 2. For detailed accuracy measurements, we calculated only the class accuracy, i.e., total correct detections of depression class testing samples against total depression class testing samples. Please note that since depression detection is a binary-target problem, the scores for detailed accuracy of positive class (depression) and binary recall will be the same.

Table 2. Measurement functions and formulas

No	Name	Function	Equation
1	Accuracy	accuracy_score()	$A(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{k=0}^{n_{samples-1}} 1(\widehat{y_k} = y_k)$ where *y* is the set of predicted pairs, $\hat{y}$ is the set of true pairs, and $n_{samples}$ is total samples.
2	Precision	precision_score()	$P(y_l, \hat{y_l}) = \frac{tp}{tp + fp}$ where *tp* is true positive and *fp* is false positive.

3	Recall	recall_score()	$R(y_l, \hat{y_l}) = \frac{tp}{tp + fn}$ where *fn* is false negative.
4	F-measure/F1	f1_score()	$F_1(y_l, \hat{y_l}) = 2 * \frac{P(y_l, \hat{y_l}) * R(y_l, \hat{y_l})}{P(y_l, \hat{y_l}) + R(y_l, \hat{y_l})}$

DATASETS

We primarily used datasets from Shen et al. [11] and Eye [12] in our experiments. These datasets were used to train and test the ML models using 10-fold CV. The texts in these two datasets were gathered from Twitter and automatically labelled as "Depression" and "Non-Depression".

The dataset by Shen et al. [11] was constructed with the additional restriction that a record would be labelled as "Depression" only if its anchor tweets satisfied the strict pattern "(I'm/ I was/ I am/ I've been) diagnosed with depression"; the record would be labelled as "Non-Depression" if the user had never posted any tweet containing the character string "depress". It should be noted that the dataset by Shen et al. [11] also contains an additional third set labelled as "Depression-candidate"—consisting of tweets that do not meet all the criteria to be labelled "Depression"—that was intended to boost the "Depression" group. However, this third set was not used in our study, because the sizes of the depression class (first set) and non-depression class (second set) are almost equal in this dataset. Including the third set in the “Depression” class would make the dataset heavily imbalanced. The dataset by Eye [12], on the other hand, only seeks the word "depression" in the tweets. If the tweet contains the word “depression”, it is labelled as "Depression", and "Non-Depression" otherwise. This dataset is highly imbalanced: depression class records are only 22% of the total records in this dataset.

To investigate whether the models trained with the above datasets could be successfully applied directly to other social media texts—besides the testing conducted with different subsets of the same dataset in 10-fold CV—we also ran the trained models on three non-Twitter based depression-class-only datasets [13-15]. These datasets could not be used to train the classifier models, since they contain only depression class samples, but at least two classes (binary) are required to train the classifier models. The first dataset, by Tanwar [13], was constructed from the diary of a 17-year old girl, Victoria, who committed suicide because of depression. In her diary, Victoria recorded her feelings from during her depression until her suicide. The second dataset, by Komati [14], was constructed from Reddit posts about depression and suicide; in this study, we employed only a part of the depression section (50000 randomly selected records from a total of 348723 records). The third dataset was constructed by Virahonda [15] from Facebook posts. Studying the performance of our models on the above datasets allowed us to test whether they are generalised enough to be applied for detecting depression by analysing texts from diverse sources. Detailed information of all the datasets used in this study can be found in Table 3.

Table 3. Datasets used in this study.

Dataset	Source	Total records	Depression records		Non-Depression records	
			Total	%	Total	%
Shen et al. [11]	Twitter	11877	6493	54.67	5384	45.33
Eye [12]	Twitter	10314	2314	22.44	8000	77.56
Tanwar [13]	Victoria’s diary	62	62	100	0	0
Komati* [14]	Reddit about depression	50000	50000	100	0	0

Virahonda [15]	Facebook	9178	9178	100	0	0

(*) randomly selected from 348723 records

RESULTS AND DISCUSSION

During preprocessing, components from the Natural Language Toolkit [62] and Peter Norvig's code for spelling correction [34] were applied. Similarly, components from scikit-learn were used to implement all the ML classifier models and to measure their performance (accuracy, precision, recall and F1) [37]. In this section, we discuss the results of applying the proposed approach to depression detection using social media texts. We take depression as the positive class and non-depression as the negative class. All experiments were conducted using the 10-fold CV method, and each experiment was repeated 10 times.

Preliminary experiments on the proposed approach

Once the preprocessing and featuring steps had been applied, preliminary experiments were run using the three best single classifiers (LR, LSVM, and MLP) from our previous study [17]. The results can be found in Table 4. The results in Table 4 are excellent—almost perfect—and, therefore, too good to be true. We suspected the possibility of overfitting, i.e., the models performed well only on the datasets they had been trained on, but would not perform as well on other datasets. Therefore, we conducted further experiments to test the performance of the trained models on three depression-class-only datasets listed in Table 3 (datasets by Tanwar, Komati, and Virahonda). The results of these experiments can be seen in Table 4 (“Accuracy on Depression-class-only Datasets” column). As predicted, the models performed poorly on the depression-class-only datasets. One of the worst results was from the LR model trained using Shen et al.’s dataset when it was applied to Tanwar’s dataset (0% accuracy). Even the best result (40.54% accuracy)—provided by LR trained using Eye’s dataset and applied to Komati’s dataset—is below the expected standards.

Table 4. Results of the preliminary experiments

Dataset	Classifier	Overall measurement (%)				Detailed Accuracy (%)		Accuracy on Depression-class-only Datasets (%)		
		Acc	Prec	Rec[3]	F1	Dep[1,3]	Non-Dep[1]	Tan[2]	Koma[2]	Vira[2]
Eye	LR	99.80	100	99.09	99.54	99.09	100	1.61	40.54	16.30
	LSVM	99.78	99.87	99.14	99.50	99.14	99.96	1.77	30.32	18.20
	MLP	99.12	99.78	96.27	97.99	96.27	99.94	9.19	31.32	31.24
	DT	99.74	99.57	99.26	99.41	99.26	99.88	13.39	48.59	31.35
Shen et al.	LR	99.77	100	99.58	99.79	99.58	100	0	16.76	13.86
	LSVM	99.80	99.94	99.68	99.81	99.68	99.93	0.32	12.09	13.67
	MLP	99.57	99.94	99.26	99.60	99.26	99.93	0	4.73	13.35
	DT	99.71	99.96	99.51	99.73	99.51	99.95	6.45	49.86	25.86

[1] Dep = Depression; Non-Dep = Non-Depression.

[2] Tan = Tanwar’s dataset; Koma = Komati’s dataset; Vira = Virahonda’s dataset; these three datasets are single-class datasets comprised only of depression records.

[3] Even though binary recall and accuracy in detecting positive (depression) classes will have the same values, both are presented here to show differences in the accuracy for depression and non-depression classes.

Since the single classifier models used in the above experiments had provided excellent results in previous studies on text detection [16,17], we suspected that the datasets used for training the models might be responsible for this problem. A simple word frequency check revealed that the words “diagnose” and “depression” always exist in all the “depression” records of Shen et al.’s pre-processed dataset. This is expected since the depression records in Shen et al.'s dataset were gathered using the strict pattern "(I’m/ I was/ I am/I’ve been) diagnosed with depression". Similarly, the depression records in Eye's dataset always contain the word “depression”. As a result, the models learnt to separate 'depression' and 'non-depression’ classes based mainly on the existence of 'diagnose' and 'depression' words in the case of Shen et al.'s dataset and the word 'depression' in the case of Eye's dataset. As a result, the models could not understand more subtle depression/non-depression patterns hidden in the text. Therefore, we deleted the word "diagnose" from Shen et al.'s pre-processed dataset and the word 'depression' from Shen et al.'s and Eye's pre-processed datasets. Following this, we conducted another set of experiments on these modified datasets, and the results can be seen in Table 5.

Table 5 Experimental results after ‘depression’ and ‘diagnose’ were deleted from the datasets

Dataset	Classifier	Overall measurement (%)				Detailed Accuracy (%)		Accuracy on Depression-class-only Datasets (%)		
		Acc	Prec	Rec[3]	F1	Dep[1,3]	Non-Dep[1]	Tan[2]	Koma[2]	Vira[2]
Eye	LR	92.61	93.32	72.21	81.38	72.21	98.51	69.19	90.40	69.65
	LSVM	91.57	84.84	76.04	80.17	76.04	96.07	60.00	85.92	70.44
	MLP	89.56	77.77	75.02	76.29	75.02	93.77	53.39	75.62	68.13
	DT	86.00	68.24	70.52	69.31	70.52	90.47	59.84	75.26	66.32
Shen et al.	LR	88.62	92.63	86.03	89.20	86.03	91.75	27.42	46.17	66.43
	LSVM	87.38	90.01	86.51	88.22	86.51	88.43	30.65	51.73	68.99
	MLP	85.13	86.65	86.07	86.35	86.07	84.01	28.39	52.70	64.30
	DT	82.23	83.46	84.20	83.82	84.20	79.86	47.90	52.62	66.93

[1] Dep = Depression; Non-Dep = Non-Depression.

[2] Tan = Tanwar’s dataset; Koma = Komati’s dataset; Vira = Virahonda’s dataset; these three datasets are single-class datasets comprised only of depression records.

[3] Even though binary recall and accuracy in detecting positive (depression) classes will have the same values, both are presented here to show differences in the accuracy for depression and non-depression classes.

A comparison of the results in Table 4 and Table 5 reveals a decrease in the overall measurements of all classifiers. The deterioration of the results is especially noticeable for the accuracy of depression records detection, which decreased by at least 13.17% for Shen et al.'s dataset and 21.25% for Eye's dataset. However, the detection accuracy significantly improved when the trained models were run on the other three depression-class-only datasets (Tanwar’s, Komati’s and Virahonda’s datasets), with the improvement ranging from 27.42% to 67.58%. It should be noted that the models were trained using datasets with both depression and non-depression classes, however, they are also being tested on three datasets with depression only classes.

These results indicate that the models trained using the modified datasets are more general and can better detect depression in the text from datasets that they were not trained on. Table 5 also reveals that the LR provided the best overall accuracy for both Shen et al.’s and Eye’s datasets; however, the results were more diverse for the remaining datasets. Similarly, results show that Shen et al.’s dataset, whose construction was stricter than Eye’s dataset, is better for depression detection on similar data, while it is less useful for the creation of a more general detection model. It should also be noted that all three classifiers that were trained using our featuring approach on Shen et al.'s dataset achieved higher scores than the baseline results in Shen et al. (85%) [11]. We also tested the difference in accuracies between each classifier for both datasets (Eye and Shen) in Table 5 using the Wilcoxon signed-rank test; these results are shown in Table 5. Given that all experiments were conducted using 10-fold CV, i.e., each process was repeated 10 times with slightly different settings of the dataset, and each experiment was repeated 10 times, these different values were compared for each classifier. It can be seen in Table 6 that the *p*-values of all pairwise comparison are smaller than the significance level (i.e., > 0.05), implying that all results are significantly different from each other.

Table 6 Wilcoxon signed-rank test results (a. Eye's dataset; b. Shen et. al.'s dataset)

a. Eye's dataset			
	SVM	**MLP**	**DT**
LR	0.00028	< 0.00001	< 0.00001
SVM		< 0.00001	< 0.00001
MLP			< 0.00001

b. Shen et al.'s dataset			
	SVM	**MLP**	**DT**
LR	0.00026	< 0.00001	< 0.00001
SVM		< 0.00001	< 0.00001
MLP			< 0.00001

Sampling or not sampling

As can be seen in Table 3, Eye's dataset is heavily imbalanced, i.e., the 'depression' records are much more numerous than the 'non-depression' records. We implemented the proposed dynamic sampling procedure on Eye's dataset to test the effectiveness of the sampling procedure, and on Shen et al.'s dataset to check the effect of applying the same procedure on a slightly imbalanced dataset. The results of these experiments can be seen in Fig. 6 and Fig. 7.

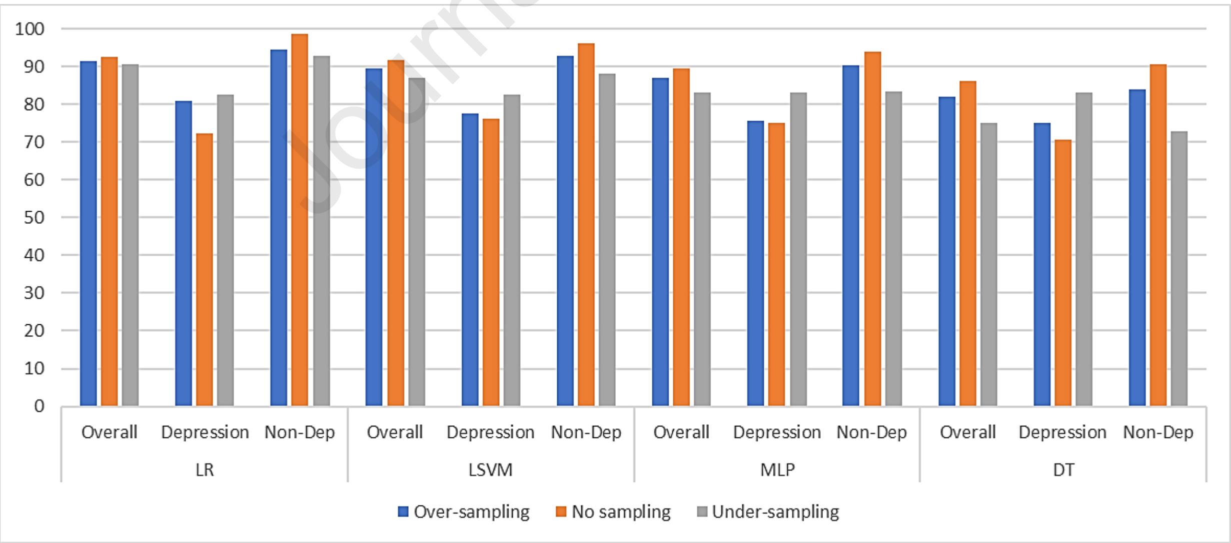


Fig. 6 Accuracy of sampling effect on Eye's dataset (heavily imbalanced)

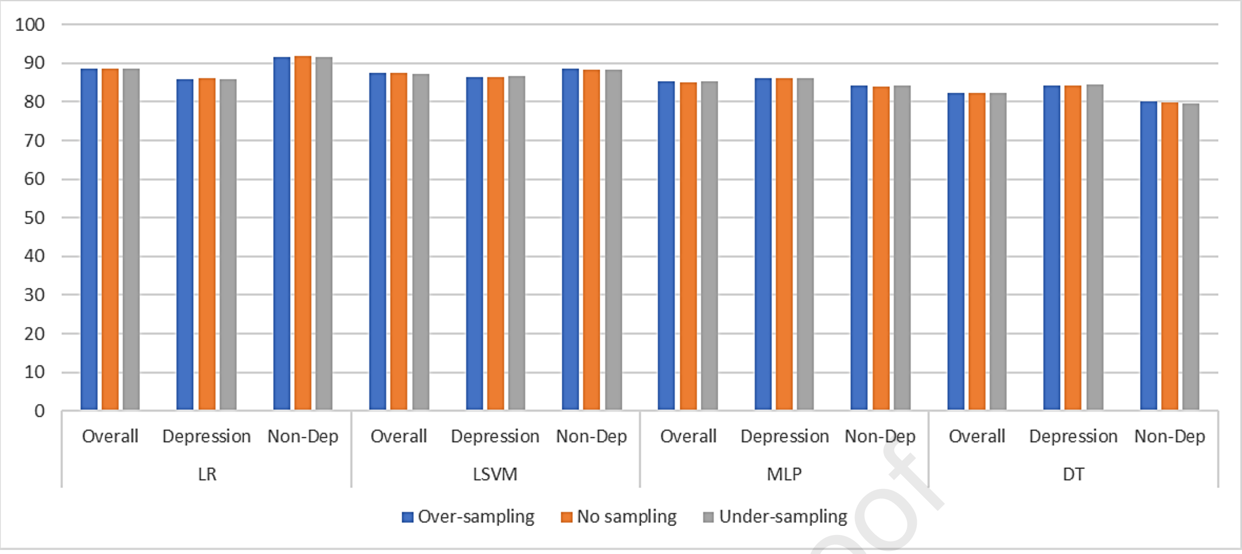


Fig. 7 Accuracy of sampling effect on Shen et al.'s dataset (slightly imbalanced)

It can be seen in Fig. 6 that both over and under-sampling improved the detection of the 'depression' class. The rates of improvement in the case of over-sampling were 0.6% with MLP, 1.5% with LSVM, 8.7% with LR, and 4.6% with DT; whereas, for under-sampling, they were 8% with MLP, 6.4% with LSVM, 10.4% with LR, and 12.5% with DT. The detection of the 'non-depression' class, however, worsened. We can conclude, from these results, that the dynamic sampling procedure can increase the detection of 'depression' class only when the dataset is imbalanced. In contrast, as can be seen in Fig. 7, dynamic sampling does not have any effect on a slightly imbalanced dataset.

The effect of dynamic sampling when models trained with Eye's and Shen et al.'s datasets were used for detecting depression in the depression-class-only datasets are shown in Fig. 8 and Fig. 9, respectively. The middle bars (i.e., no sampling) correspond to the scores in Table 5.

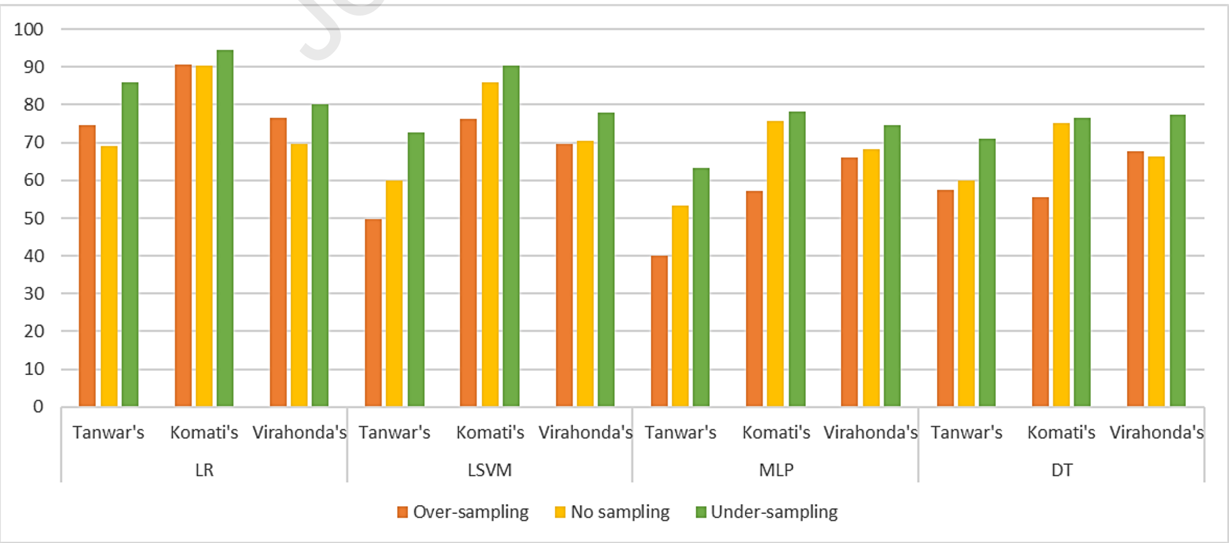


Fig. 8 Accuracy of sampling effect of Eye's trained models on depression-class-only datasets

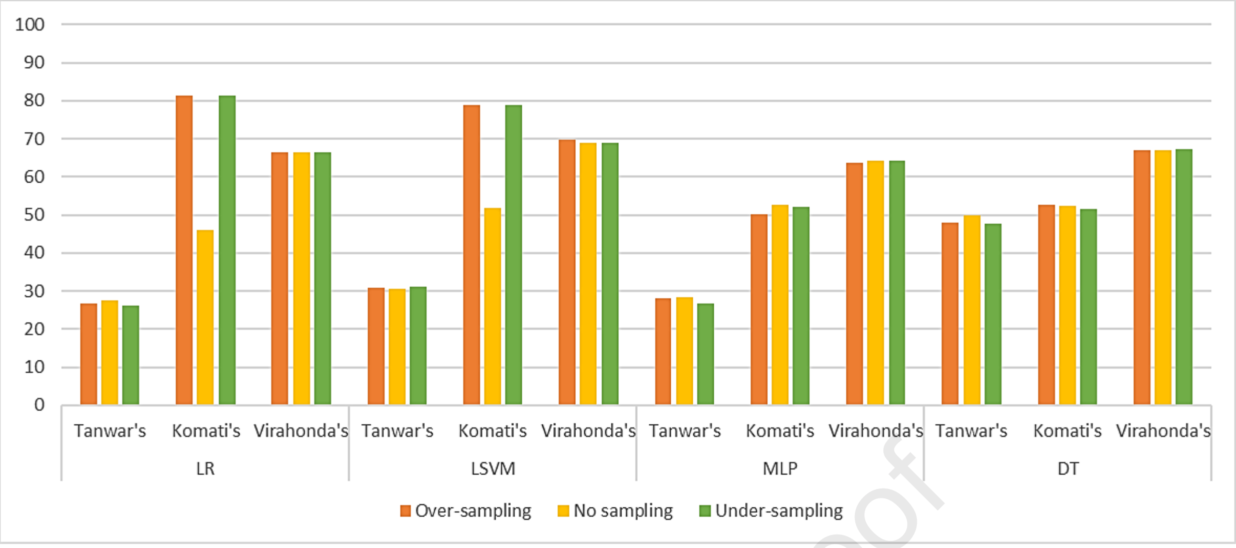


Fig. 9 Accuracy of sampling effect of Shen et al.'s trained models on depression-class-only datasets

We can see in Fig. 8 that all the models trained using Eye's dataset improved the accuracy when combined with under-sampling as opposed to when dynamic sampling was not applied. However, this effect was not always replicated in the case of over-sampling; only the LR model improved accuracy with over-sampling compared to no sampling. We suspect that, with over-sampling, the training records are duplicated randomly in the case of the smaller class (the 'depression' class in our case), which increases the population of the smaller class but not its diversity. This condition makes the trained models more exclusive to the trained dataset, but not the remaining datasets. In the case of models trained with Shen et al.'s dataset (see Fig. 9), as can be predicted from the results in Fig. 7, dynamic sampling did not have a significant impact on the detection accuracy. However, surprisingly, both sampling methods significantly improved the detection accuracy in the case of Komati's depression dataset with the LR and LSVM models. Finally, based on the results in Fig. 8 and Fig. 9, we can conclude that the LR performed the best with both sampling methods, and that dynamic under-sampling increased the accuracy of the depression detection.

Experiments on ensemble classifiers

Our last set of experiments were conducted to test the performance of our method on several classifier ensembles, namely AB, BP, GB, and RF. These ensemble models had performed well in previous studies on text analysis (e.g. see [17]). Based on the results in the previous subsection, these experiments were carried out together with dynamic under-sampling. The results of the experiments on Shen et al.'s and Eye's datasets and the depression-class-only datasets can be seen in Fig. 10 and Fig. 11, respectively.

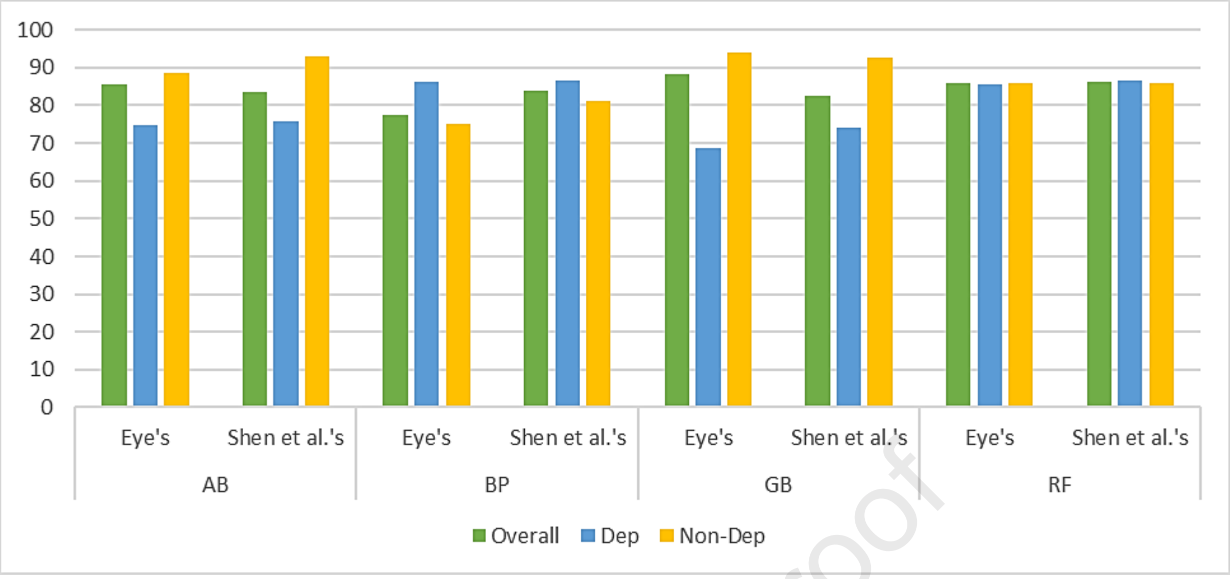


Fig. 10 Accuracy of several ensemble classifiers with under-sampling setting for Eye's and Shen et al.'s datasets

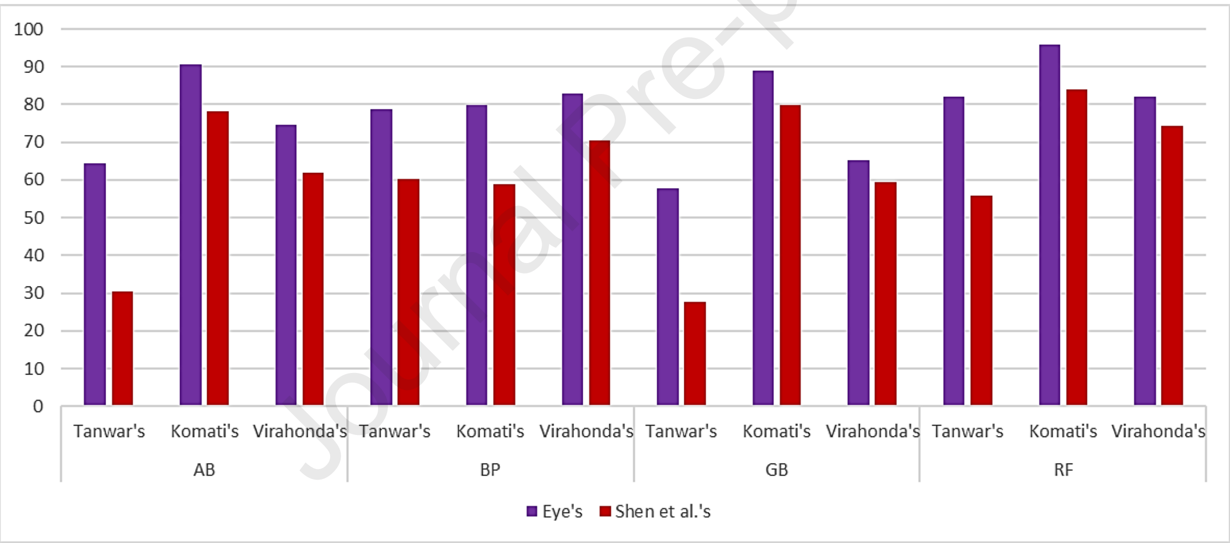


Fig. 11 Accuracy of ensemble models that are trained using Eye's and Shen et al.'s datasets to be tested on Tanwar's, Komati's and Virahonda's datasets

The results in Fig. 10 indicate that, among the classifier ensembles, RF provided the most balanced accuracy for both depression and non-depression classes in both the training datasets. However, these results were not better than those provided by the single classifier LR. While we are unable to pinpoint the reason why the results were not better than LR (numerous factors could influence the results of an ensemble), we believe the base classifiers used by the ensembles might be significantly impacting their results. As discussed above, BP, AB, and RF use DT as their base classifier. The results in Table 5 show DT to be the worst classifier among those tested, and the performance of the DT-based ensembles was similar to DT. Regarding the performance of GB, which uses a single regression tree for classification/prediction in the vein of LR, we suspect the boosting process in GB is less suitable for depression detection than LR. On the other hand, when tested on the three depression-class-only datasets (see the results in Fig. 11), RF provided better results than all single and ensemble classifiers.

Therefore, we can conclude that the RF model is better at depression detection using general text than other classifiers.

CONCLUSION

Depression is the most prevalent mental disorder and the main cause behind more than two-thirds of suicides every year. Unfortunately, many cases go untreated because of failure to detect as well as self-denial. Several studies agree that, given the exponential increase in social media usage, social media messages can be used as a valuable source for monitoring several mental health issues, including depression.

In this paper, we proved that our generalised approach using ML methods and social media texts can be effectively used to detect signs of depression. Our ML models proved effective even when trained with texts that did not contain the words 'depression' or 'diagnosis'—social media messages by those suffering from depression rarely include such words. It is important to note that the approach presented in this paper performed well even when tested on datasets that were unrelated to the training datasets. This contrasts with most studies in the literature that use portions of the same dataset for training and testing. Our results also indicate that using less strictly constructed datasets can be more beneficial than more strictly constructed datasets, especially when the models would be used for detecting depression in messages from a variety of sources.

It should be noted that the approach presented in this paper uses supervised ML classifiers, and therefore, the approach is limited to using labelled datasets for training the classifiers. Overcoming this limitation by including unsupervised classifiers is a potential future research area.

An additional observation is that, in a heavily imbalanced dataset, dynamic sampling could increase the accuracy of the less-populous class but, at the same time, decrease the accuracy of the more populous class. However, this can be beneficial if the goal is to detect depression and the less-populous class is the depression class (which is the case in this study).

REFERENCES

1. Priya A, Garg S, Tigga NP (2020) Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science 167:1258-1267

2. Srimadhur NS, Lalitha S (2020) An End-to-End Model for Detection and Assessment of Depression Levels using Speech. Procedia Computer Science 171:12-21

3. Alsagri HS, Ykhlef M (2020) Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features. IEICE Transactions on Information and Systems E103.D (8):1825-1832. doi:10.1587/transinf.2020EDP7023

4. Nikolin S, Tan YY, Schwaab A, Moffa A, Loo CK, Martin D (2021) An investigation of working memory deficits in depression using the n-back task: A systematic review and meta-analysis. Journal of Affective Disorders 284:1-8. doi:https://doi.org/10.1016/j.jad.2021.01.084

5. Yadav S, Kaim T, Gupta S, Bharti U, Priyadarshi P Predicting depression from routine survey data using machine learning. In: Proceedings of 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, Dec 18-19, 2020. pp 163-168. doi:10.1109/ICACCCN51052.2020.9362738

6. Hassan AU, Hussain J, Hussain M, Sadiq M, Lee S Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In: Proceedings of 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, South Korea, 18-20 Oct. 2017. pp 138-140. doi:10.1109/ICTC.2017.8190959

7. Edo-Osagie O, De La Iglesia B, Lake I, Edeghere O (2020) A scoping review of the use of Twitter for public health research. Computers in Biology and Medicine 122:103770. doi:10.1016/j.compbiomed.2020.103770

8. Mishra V, Garg T (2018) A systematic study on predicting depression using text analytic. Journal of Fundamental and Applied Sciences 10 (2):293-307. doi:10.4314/jfas.v10i2.21

9. Kim J, Lee J, Park E, Han J (2020) A deep learning model for detecting mental illness from user content on social media. Scientific reports 10 (1):11846-11846. doi:10.1038/s41598-020-68764-y

10. Mahnken K (2021) Survey: More Young People Are Depressed During the Pandemic. But They May Be Using Social Media to Cope. The 74 million. https://www.the74million.org/survey-more-young-people-are-depressed-during-the-pandemic-but-they-may-be-using-social-media-to-cope/. Accessed 28 April 2021

11. Shen G, Jia J, Nie L, Feng F, Zhang C, Hu T, Chua T-S, Zhu W Depression detection via harvesting social media: A multimodal dictionary learning solution In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19-25 August 2017. pp 3838-3844. doi:https://doi.org/10.24963/ijcai.2017/536

12. Eye BB (2020) Depression Analysis. 1 edn., Kaggle

13. Tanwar R (2020) Victoria Suicide Data. Kaggle,

14. Komati N (2020) r/SuicideWatch and r/depression posts from Reddit. Kaggle,

15. Virahonda S (2020) Depression and anxiety comments. 1 edn. Kaggle,

16. Budhi GS, Chiong R, Wang Z (2021) Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. Multimedia Tools and Applications. doi:https://doi.org/10.1007/s11042-020-10299-5

17. Budhi GS, Chiong R, Pranata I, Hu Z (2021) Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis. Archives of Computational Methods in Engineering. doi:https://doi.org/10.1007/s11831-020-09464-8

18. Budhi GS, Chiong R, Pranata I, Hu Z Predicting rating polarity through automatic classification of review texts. In: Proceedings of the 2017 IEEE Conference on Big Data and Analytics (ICBDA), Kuching, Malaysia, November 16-17, 2017. pp 19-24. doi:10.1109/ICBDAA.2017.8284101

19. Sato JR, Moll J, Green S, Deakin JF, Thomaz CE, Zahn R (2015) Machine learning algorithm accurately detects fMRI signature of vulnerability to major depression. Psychiatry Res 233 (2):289-291. doi:10.1016/j.pscychresns.2015.07.001

20. Kumar P, Garg S, Garg A (2020) Assessment of anxiety, depression and stress using machine learning models. Procedia Computer Science 171:1989-1998. doi:https://doi.org/10.1016/j.procs.2020.04.213

21. Ojeme B, Mbogho A (2016) Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders. Procedia Computer Science 96:1294-1303. doi:10.1016/j.procs.2016.08.174

22. Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum D, Rizzo S, Morency L-P The Distress Analysis Interview Corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, May 26-31, 2014. European Language Resources Association (ELRA), pp 3123-3128

23. Samareh A, Jin Y, Wang Z, Chang X, Huang S (2018) Detect depression from communication: How computer vision, signal processing, and sentiment analysis join forces. IISE Transactions on Healthcare Systems Engineering 8 (3):196-208. doi:10.1080/24725579.2018.1496494

24. Chen Y, Zhou B, Zhang W, Gong W, Sun G Sentiment Analysis Based on Deep Learning and Its Application in Screening for Perinatal Depression. In: Proceedings of 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), 2018. pp 451-456. doi:10.1109/dsc.2018.00073

25. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A (2018) Depression detection from social network data using machine learning techniques. Health Inf Sci Syst 6 (1):8. doi:10.1007/s13755-018-0046-0

26. Burdisso SG, Errecalde M, Montes-y-Gómez M (2019) A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications 133:182-197. doi:10.1016/j.eswa.2019.05.023

27. Fatima I, Abbasi BUD, Khan S, Al-Saeed M, Ahmad HF, Mumtaz R (2019) Prediction of postpartum depression using machine learning techniques from social media text. Expert Systems 36 (4). doi:10.1111/exsy.12409

28. Chiong R, Budhi GS, Dhakal S (2021) Combining sentiment lexicons and content-based features for depression detection. IEEE Intelligent Systems 36 (6)

29. Filho EMS, Veiga Rey HC, Frajtag RM, Arrowsmith Cook DM, Dalbonio de Carvalho LN, Pinho Ribeiro AL, Amaral J (2021) Can machine learning be useful as a screening tool for depression in primary care? J Psychiatr Res 132:1-6. doi:10.1016/j.jpsychires.2020.09.025

30. Jothi N, Husain W, Rashid NA (2020) Predicting generalized anxiety disorder among women using Shapley value. J Infect Public Health 14 (1):103-108. doi:10.1016/j.jiph.2020.02.042

31. Jung H, Park HA, Song TM (2017) Ontology-Based Approach to Social Data Sentiment Analysis: Detection of Adolescent Depression Signals. J Med Internet Res 19 (7):e259. doi:10.2196/jmir.7452

32. Sutter B, Chiong R, Budhi GS, Dhakal S Predicting psychological distress from ecological factors: A machine learning approach. In: Proceedings of the 34th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE 2021), Kuala Lumpur, Malaysia, July 2021.

33. Lin C, Hu P, Su H, Li S, Mei J, Zhou J, Leung H (2020) SenseMood: Depression Detection on Social Media. In: Proceedings of the 2020 International Conference on Multimedia Retrieval. Association for Computing Machinery, pp 407–411. doi:10.1145/3372278.3391932

34. Norvig P (2016) How to write a spelling corrector. https://norvig.com/spell-correct.html. Accessed June 01 2018

35. Etaiwi W, Naymat G (2017) The impact of applying different preprocessing steps on review spam detection. Procedia Computer Science 113:273-279

36. Deng X, Li Y, Weng J, Zhang J (2019) Feature selection for text classification: A review. Multimedia Tools and Applications 78 (3):3797-3816. doi:10.1007/s11042-018-6083-5

37. Hu Z, Chiong R, Pranata I, Bao Y, Lin Y (2019) Malicious web domain identification using online credibility and performance data by considering the class imbalance issue. Industrial Management & Data Systems 119 (3):676-696. doi:10.1108/IMDS-02-2018-0072

38. Nelder JA, Wedderburn RWM (1972) Generalized Linear Models. Journal of the Royal Statistical Society Series A (General) 135 (3):370-384. doi:10.2307/2344614

39. Hastie T, Tibshirani R (1990) Generalized Additive Models. Chapman and Hall/CRC, United Kingdom

40. Dunteman GH, Ho M-HR (2011) Generalized Linear Models. In: An Introduction to Generalized Linear Models. SAGE Publications, Inc., pp 2-6

41. Dobson AJ, Barnett AG (2008) An Introduction to Generalized Linear Models. 3rd edn. CRC Press, Boca Raton

42. Menard S (2010) Logistic Regression: From Introductory to Advanced Concepts and Applications. SAGE, Los Angeles

43. Chiong R, Fan Z, Hu Z, Chiong F (2021) Using an improved relative error support vector machine for body fat prediction. Computer Methods and Programs in Biomedicine 198:105749

44. Campbell C, Ying Y (2011) Learning with Support Vector Machines. Morgan & Claypool,

45. Chang CC, Lin CJ (2011) LIBSVM: A library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2 (3):1-27. doi:10.1145/1961189.1961199

46. Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, Arshad H (2018) State-of-the-art in artificial neural network applications: A survey. Heliyon 4 (11):e00938. doi:10.1016/j.heliyon.2018.e00938

47. Abiodun OI, Kiru MU, Jantan A, Omolara AE, Dada KV, Umar AM, Linus OU, Arshad H, Kazaure AA, Gana U (2019) Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition. IEEE Access 7:158820-158846. doi:10.1109/access.2019.2945545

48. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol 1. MIT Press, pp 318-362

49. Kingma DP, Ba J Adam: A method for stochastic optimization. In: Proceedings of International Conference on Learning Representations, San Diego, US, May 7-9, 2015. pp 1-15

50. Quinlan JR (1986) Induction of decision trees. Machine Learning 1 (1):81-106. doi:10.1007/bf00116251

51. Hunt EB, Marin J, Stone PJ (1966) Experiments in induction. Academic Press, New York

52. Breiman L (2001) Random forests. Machine Learning 45 (1):5-32. doi:10.1023/a:1010933404324

53. Freund Y, Schapire RE (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 55 (1):119-139. doi:https://doi.org/10.1006/jcss.1997.1504

54. Zhu J, Zou H, Rosset S, Hastie T (2009) Multi-class AdaBoost. Statistics and Its Interface 2:349-360

55. Breiman L (1996) Bagging predictors. Machine Learning 24 (2):123-140. doi:10.1007/bf00058655

56. Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29 (5):1189-1232

57. Ren Q, Li M, Han S (2019) Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives. Big Earth Data 3 (1):8-25. doi:10.1080/20964471.2019.1572452

58. Xiong T, Bao Y, Hu Z, Chiong R (2015) Forecasting interval time series using a fully complex-valued RBF neural network with DPSO and PSO algorithms. Information Sciences 305:77-92. doi:10.1016/j.ins.2015.01.029
59. Primartha R, Adhi Tama B, Arliansyah A, Januar Miraswan K (2019) Decision tree combined with PSO-based feature selection for sentiment analysis. Journal of Physics: Conference Series 1196. doi:10.1088/1742-6596/1196/1/012018
60. Nguyen T-L, Kavuri S, Lee M, Hwang SO (2018) A fuzzy convolutional neural network for text sentiment analysis. Journal of Intelligent & Fuzzy Systems 35 (6):6025-6034. doi:10.3233/jifs-169843
61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12 (85):2825-2830
62. Bird S, Klein E, Loper E (2009) Natural Language Processing with Python. O'Reilly Media, Inc., USA

Highlights

- Depression is among the most prevalent mental disorder that can lead to suicide.
- Due to self-denial, depression can remain untreated and can aggravate the condition.
- Social media texts are useful for monitoring depression.
- Textual-based featuring methods alongside machine learning classifiers were tested.
- Results show that the tested approach performs well across different datasets.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: