Detecting Signs of Depression Using Social Media Texts Through an Ensemble of Ensemble Classifiers

Raymond Chiong [a,*], Gregorius Satia Budhi [b,*], Erik Cambria [c], *IEEE Fellow*

Abstract—Artificial intelligence-based machine learning models have been widely used to explore and address various mental health-related problems in recent years, including depression. In this study, we present an ensemble approach to complement the 90 unique input features that we proposed in a previous study on depression detection using social media texts. Our proposed Ensemble of Ensemble Classifiers (EECs) combines many ensemble models, including Bagging Predictors, Random Forests, Adaptive Boosting and Gradient Boosting, as inner ensembles. These inner ensembles are arranged in a parallel fashion, where each of them is trained using different subsets of data sampled from the training data via bootstrap sampling. After the models are trained, during the testing phase, the results of all inner ensembles are processed using two methods—majority vote or class priority threshold—to get the final result as an output. From the experiments, we find that EECs are accurate in detecting signs of depression in social media users by analysing their posts in social media platforms such as Twitter. Our approach outperforms other ensemble methods on the public datasets we used. Moreover, if set correctly, the parameters of EECs can further improve the performance of the proposed ensemble in detecting signs of depression.

Index Terms—Depression detection, social media, ensemble models, machine learning

1 Introduction

Depression is an aversive state that usually accompanies negative thinking about oneself, current surroundings and what is likely to happen in the future. This state is sometimes characterised by lower levels of interest and pleasure, known as anhedonia. The core symptoms of depression are frequent sadness, a lack of energy and a reduced ability to enjoy things [1]. However, because of poor recognition and denial of this state, depression can remain undiagnosed or untreated [2]. For example, three out of four persons with severe mental disorders are often untreated, thus making the problem acute [3]. A lack of diagnosis can further aggravate the condition, which can reach life-threatening severity (e.g., suicide attempt [4]), if the person enters a cycle of negative affect, cognition and behaviour and can no longer respond to positive influences [1, 2]. This situation can lead to a reduced quality of life and, in acute cases, an inability to maintain employment [5, 6]; worse, depression is the cause of more than two-thirds of suicides [7]. To further understand and address depression, researchers from different disciplines have begun using computational models for depression detection and treatment; in particular, an emerging field known as computational psychiatry [8] focuses on applying computational approaches to study mental illness and other mental health problems [1].

Computational psychiatry relies on both data-driven and theory-driven approaches [9]: the data-driven approach implements machine learning (ML) to process high-dimensional data to achieve better classification of disease, predict treatment outcomes or improve treatment selection, while the theory-driven approach applies models to instantiate prior knowledge of, or hypotheses about, the underlying mechanisms at multiple levels of analysis and abstraction [9]. The data-driven approach is typically used when data samples from previous cases exist in large quantities, whereas the theory-driven approach is often applied in more traditional settings that involve developing hypotheses and collecting data to test the hypotheses.

Furthermore, several studies on mental health have shown that social media posts can be utilised to monitor health issues and trends [10]. People often express their feelings and share their thoughts on social media platforms [11, 12] before seeing health professionals [12]. These behaviours allow researchers to investigate aspects of psychological concerns regarding human behaviour in such contexts. Some studies dedicated to depression have found that tweets posted by individuals with a major depressive disorder can be used to predict if they are likely to suffer a future depression episode [2]. In this study, we propose an improved ML ensemble approach, the Ensemble of Ensemble Classifiers (EECs). While other ensemble models utilise a single type of classifier as their base classifier/detector, we implement ensembles as base detectors. We call these inside detectors 'inner ensembles'. Inner ensembles can be sourced from the same type of ensemble model or a combination of multi-type ensemble models.

- [a] *School of Information & Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia. E-mail: Raymond.Chiong@newcastle.edu.au*
- [b] *Informatics Department, Petra Christian University, Surabaya 60236, Indonesia. E-mail: greg@petra.ac.id*
- [c] *School of Computer Science and Engineering, Nanyang Technological University, 639798 Singapore E-mail: cambria@ntu.edu.sg*

By combining ensembles, our model is the continuation of our previous study on depression detection [13, 14], which outperforms the results of other ensemble models. While we achieved excellent results in that study, we perceived a way to improve the detection, especially the detection of depression class. As in the previous study, we utilise two labelled datasets gathered from Twitter [15, 16]. The records in these datasets are labelled 'depression' or 'non-depression' Twitter messages. Instead of investigating how Twitter text is preprocessed or extracted, we use the preprocessing and features extracted from previous research. This study contributes to ML research areas, such as ensemble models and natural language processing. It also contributes to improving mental health by proposing a model that could effectively detect depression via social media, especially when users are not aware of their depression.

The rest of this paper is organised as follows. In the next section, we discuss related work, and then explain the design of our proposed model, the EECs. After that, we discuss how we conducted the experiments to test the performance of EECs. Next, we discuss the results, and finally, we conclude the study and outline the possibilities for future work.

2 Related work

Computational psychiatry has emerged as a hot research topic. This field has become prominent following advances in computational neuroscience and ML, and the scientific understanding of psychiatric disorders, including depression and suicide tendency [17]. The study encompasses two approaches. The first is data-driven, agnostic data analysis broadly using ML and standard statistical methods. This approach has been applied to several clinically relevant problems, such as automatic diagnosis, prediction of treatment outcomes and treatment selection [9]. The second draws on theory-driven models that mathematically specify interpretable relations between variables, including meaningful hidden variables [9]. Several kinds of research in this study focus on applying computational neuroscience to the study of mental disorders. Research on mental disorders such as depression detection involves medical data, such as results of depression questionnaires (such as DASS21 and DASS42) [18, 19], clinical criteria for depression as defined in DSM-5 and ICD-10 [20], as well as EEG [21] and fMRI data [22, 23]. Data from clinical interviews, using systems such as Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ [24]) [7, 25], have also been collected. Data from DAIC-WOZ include videos, speeches and text transcriptions from distressed and non-distressed participants.

Other researchers have followed another course and sought to detect mental disorders such as depression and suicidal thoughts through facial expressions [26] or text messages on social media platforms, such as Twitter, Facebook, Reddit and WeChat [2, 10, 15, 27-31]. The hope is that social media texts may help detect mental disorders even when those afflicted are unaware of their illness or have not been able to obtain help. Most research on mental disorder detection via social media messages follows either a text-based featuring approach or a personal descriptive-based featuring approach. Textual-based featuring focuses on the linguistic features of the social media text [2, 10, 13, 27, 29-31]. In contrast, the descriptive-based featuring approach focuses on descriptions of the subject or patient [15, 20, 32-35]. These features then become the input of the detection models. Most depression detection systems 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) and K-Nearest Neighbours (KNN). They have also implemented ensemble models, such as Adaptive Boosting (AB), Random Forest (RF), Gradient Boosting (GB) and Bagging Predictors (BP) [2, 10, 13, 14, 18, 20, 22, 30, 32, 33, 35]. Deep learning methods, such as the Long Short-Term Memory (LSTM) [28] and Convolutional Neural Network (CNN) [27, 36], have also been used. Additionally, several studies have also constructed custom detectors [15, 30, 34]. An overview of related work over the past five years (2017–2021) is provided in Table 1.

In our previous studies on depression detection in social media, our focus was on how to create features for ML from raw text messages in social media such as Twitter. The first study [14] was designed to utilise ML methods combined with textual-based featuring. These features were created using text preprocessing and featuring methods, such as tokenisation, stop word removal, detection of negation words, correction of elongation words, parts of speech (POS) lemmatisation, bag-of-words (BOW) and n-gram words. We then applied ML classifiers—both single and ensemble models—that are widely used in solving prediction problems. In the second study [13], we used a different approach to extract input features from the text, introducing 90 unique features to be processed by ML. This study combined feature extraction using sentiment lexicons and textual content–based features from social media texts. Two sentiment lexicons [37, 38] and the characteristics of the Twitter textual content (e.g. the number of words, sentences, questions, exclamations, POS tags, linguistic traits and readability scores) were used to build the features.

We do not compare our studies to other studies in Table 1, as the datasets used are different, and the formulas for the measurements may also differ. However, in general, our studies achieved good results, especially the second study [13], which consistently gave the best results, with all measurement scores above 95%. Therefore, instead of working on another feature-processing approach in the current study, we designed an ensemble model, drawing on feature extraction methods designed previously in Chiong et al. [13]. Unlike other ensembles that are the combination of classifiers, the EECs combine ensemble models together in unity to detect signs of depression.

TABLE 1 OVERVIEW OF RELATED WORK (2017–2021)

Author	Year	Feature type	Dataset source	Best result[1]
Depression Detection on Social Media				
Shen et al. [15]	2017	Descriptive-based	Twitter (Public/Pu, [15])	Acc: 85%; Pre: 85%; Rec: 85%; F1: 85%
Hassan et al. [10]	2017	Textual-based	Twitter (Private/Pr)	Acc: 91%; Pre: 83%; Rec: 79%
Chen et al. [28]	2018	Textual-based	Survey and WeChat (Pr)	Present the results in several graphs
Islam et Al. [29]	2018	Textual-based	Facebook (Pr)	Pre: 59%; Rec: 97%; F1: 73%
Burdisso et al. [30]	2019	Textual-based	Reddit (Pr)	Pre: 63%; Rec: 60%; F1: 61%
Fatima et al. [31]	2019	Textual-based	Reddit (Pr)	Acc: 91.63%; Pre: 91.83%; Rec: 91.85%
Lin et al. [36]	2020	Visual- and textual-based	Twitter (Pu, [15]) and Images	Acc: 88.4%; Pre: 90.3%; Rec: 87%; F1: 93.6%
Alsagri & Mourad [2]	2020	Textual-based	Twitter (Pr)	Acc: 82.5%; Pre: 73.91%; Rec: 85%; F1: 79.06%; AUC: 0.78
Kim et al. [27]	2020	Textual-based	Reddit (Pu, [27])	Acc: 75.13%; Pre: 89.1%; Rec: 71.75%; F1: 79.49%
Chiong et al. [14]	2021	Textual-based	Twitter (Pu,[15, 16]); Facebook (Pu, [39]); Reddit (Pr); e-Diary (Pu, [40])	Twitter [15] (Acc: 88.62%; Pre: 92.63%; Rec: 86.03%; F1: 89.2%); Twitter [16] (Acc: 92.61%; Pre: 93.32%; Rec: 72.21%; F1: 81.38%)
Chiong et al. [13]	2021	Hybrid sentiment-based and textual content–based	Twitter (Pu,[15, 16])	Twitter [15] (Acc: 98.05%; Pre: 97.87%; Rec: 98.59%; F1: 98.22%); Twitter [16] (Acc: 98.05%; Pre: 95.11%; Rec: 96.30%; F1: 95.69%)
Depression Detection from Other Sources (Non–Social Media)				
Jung et al. [34]	2017	Textual-based/ Ontology	35 FAQs about depression from multi sources	Acc: 75% ; Pre: 76.1%
Samareh et al. [25]	2018	Audio-, video- and textual-based	DAIC-WOZ[1] (Pu, [24])	RMSE: 5.12; MAE: 4.12
Priya et al. [18]	2020	Descriptive-based	DASS-21[1] (Pr)	Acc: 85.5%; Pre: 82.2%; Rec: 85%; F1: 83.6
Kumar et al. [19]	2020	Descriptive-based	DASS-42[1] (Pr)	Acc: 96%; Pre: 96%; Rec: 96%; F1: 96%; AUC: 0.99
Srimadhur & Lalitha [7]	2020	Spectrogram & End-to-end	DAIC-WOZ (Pu, [24])	Pre: 65%; Rec: 92%; F1: 76%
Jothi et al. [33]	2020	Descriptive-based	Online survey (Pr)	Acc: 95.7%; Rec: 97.5%; Spe: 86.3%
Filho et al. [32]	2021	Descriptive-based	Patients clinical evaluation (Pr)	Acc: 89%
Chen et al. [26]	2021	Sequences of facial region frames	Video (Pu, [41])	MAE: 6.16; RMSE: 8.13

[1] Acc = Accuracy; Pre = Precision; Rec = Recall; Spe = Specificity; F1 = F-measure; AUC = Area Under the Receiver-Characteristic-Operator (ROC) Curve; MAE = Mean Absolute Error; RMSE = Root Mean Square Error; DASS = Depression, Anxiety and Stress Scale questionnaire; DAIC-WOZ = The Distress Analysis Interview Corpus-Wizard of Oz.

3 ENSEMBLE OF ENSEMBLE CLASSIFIERS (EECS)

The EEC is a novel ensemble model developed to improve the performance of other ensemble models. While it is possible to implement the model for other problems, we designed it specifically to improve our previous work on depression detection in social media [12, 13]. The design is depicted in Fig. 1.

Fig. 1 depicts two processes: the training process (the full head arrows ➔) and the testing process (the line head arrows →). All processes begin with inputting the training or testing sets and the parameter settings of the EECs (see hollow head arrows -▷).

For training purposes, we first need to choose the type of inner ensembles and the total number of inner ensembles used. The type of these inner ensembles could be any ensemble classifier, i.e. AB, GB, and BP. It is possible to choose the percentage of each inner ensemble used in the process. However, for ease, the default setting sets the inner ensembles to be created in equal numbers; for example, suppose a total of ten ensembles across three types (AB,

GB, BP) are used, then the equal setting will create four ABs, three GBs and three BPs. Changing the default setting for the number of inner ensembles is suggested if the user knows the precise purpose for the setting. For example, if the user knows that a particular ensemble type is suitable for the problem, they can have more inner ensembles of this type than other ensemble types.

The inner ensemble is structured as an array of ensembles that work in parallel with no particular order. After creating the inner ensembles following the setting, for each inner ensemble model is assigned a subset of training samples based on the bootstrap sampling process. The bootstrap sampling method draws random sample data repeatedly with replacement from the source, based on the percentage setting [41]. For example, if the setting is 0.25 (25%) and the total training sample data is 1000 records, then each inner ensemble will be trained using 250 records that are taken randomly with replacement from the training sample data. After the training process for each inner ensemble model, the weights and parameters of all inner ensemble models are saved in a file.

The testing process is straightforward. After inputting the testing samples and loading the previously trained inner ensemble models, all samples are run with the inner ensemble models. The final output is determined by choosing one of the following two processes: the Majority Vote or the Priority Threshold. The Majority Vote chooses the majority class outputs from the results of the inner ensembles. If the majority votes of all classes are equal, the final result is the smallest class in the corpus (the class with the fewest number of sample records in the corpus). The Priority Threshold idea is the opposite of Majority Vote. It provides a final result based on the priority chosen in the setting (a.k.a minority could win if prioritised). For example, if the positive class is chosen, then if the positive class outputs from the inner ensembles are the same as or more than the threshold, the final output is a positive class. The priority threshold can be set from absolute, which needs only one inner ensemble to pick the chosen class, to the maximum threshold before it becomes a majority vote (i.e., more than 50% + 1 in a binary classification problem). The user can then choose to show the actual vote percentage of each class if needed.

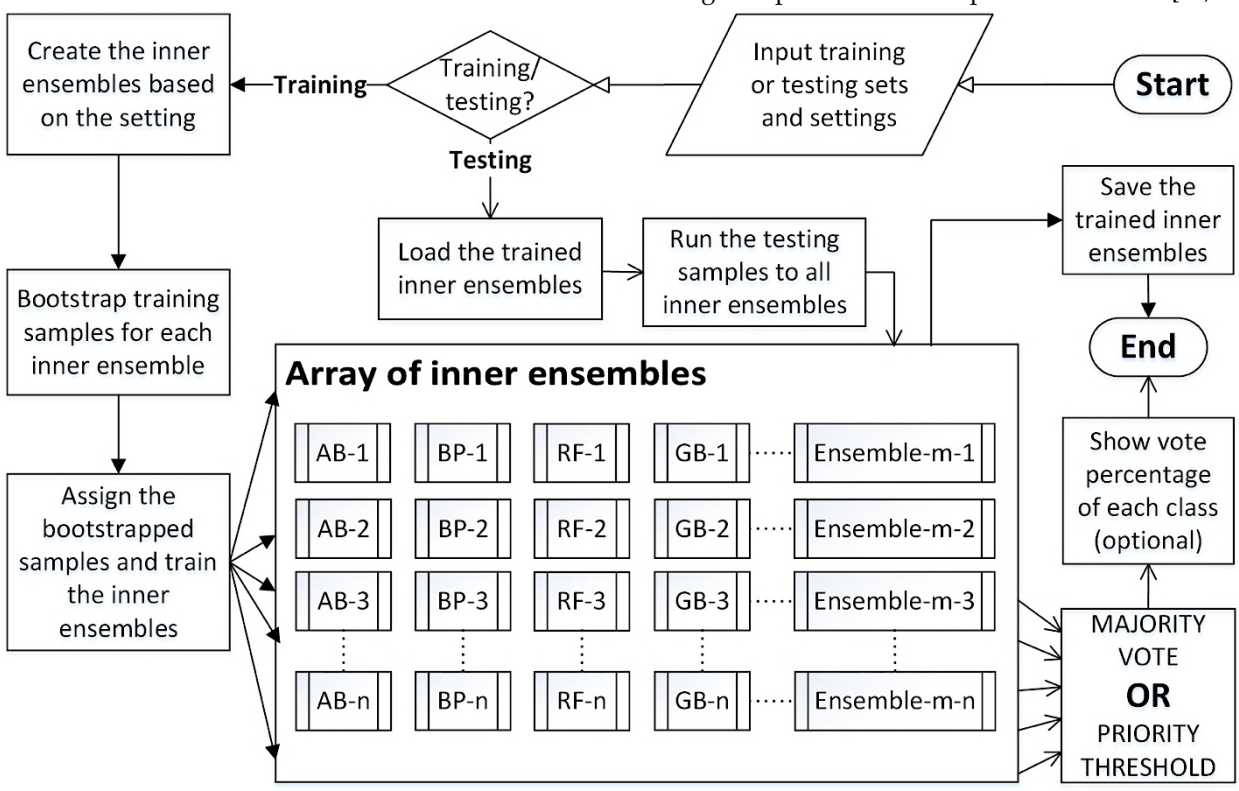

Fig. 1 Design of the EECs

4 Experimental Setup

4.1 Datasets

Intuitively, the EECs model can be applied to many problems, similar to other ML ensemble models. However, here, we focus on applying the EECs for depression detection in social media. Therefore, we conducted experiments using two public Twitter depression datasets [15, 16] that we have used previously [13, 14] (see Table 2 for additional details).

The two depression datasets in Table 2, comprising Twitter posts that have automatically been labelled either 'depression' or 'non-depression', were used to train and test the EECs using 10-fold cross-validation (CV).

Shen et al.'s dataset [15] was constructed with the restriction that a record would be labelled 'Depression' only if the anchor tweets satisfied the exact pattern '(I'm / I was

/ I am / I've been) diagnosed with depression'. The record is labelled 'non-depression' if the user has never posted a tweet containing the character string 'depress'. Eye's dataset [16], in contrast, is less restrictive and was built by seeking the word 'depression' in tweets; any tweet containing the word 'depression' was labelled 'Depression', and 'non-depression' otherwise. Eye's dataset is highly imbalanced; depression class records account for only 22% of the total records. In contrast, Shen et al.'s dataset is slightly imbalanced, with depression records slightly outnumbering non-depression records.

TABLE 2 STATISTICS FOR THE PUBLIC DATASETS USED IN THIS STUDY

Dataset	Source	Total records	Depression records		Non-depression records	
			Total	%	Total	%
Eye [16]	Twitter	10314	2314	22.44	8000	77.56
Shen et al. [15]	Twitter	11877	6493	54.67	5384	45.33

4.2 Framework for testing

To evaluate the EECs, we implemented the comparison framework shown in Fig. 2. We used this framework previously to investigate classifiers for sentiment polarity detection of customer reviews [42] and to identify the best models for fake review detection [43, 44]. In this study, we modified the framework so that it could be applied to test the EECs for depression detection. We used this framework to test and compare the performance of different settings of the EECs using the Twitter depression datasets depicted in Table 2.

The input features in this study follow the approach in [13], which achieved good results. These 90 input features were categorised into two groups (see Table 3). The first group comprises features that have been defined based on two sentiment lexicons: SentiWordNet by Baccianella et al. [37] and SenticNet by Cambria et al. [38]. Group A consists of nine features created using SentiWordNet, whereas group B consists of 13 features extracted using SenticNet.

The second group consists of 68 features defined in one of our previous studies [44]. These features were extracted from the characteristics of tweets and categorised into four groups (C, D, E and F) as follows. The features in group C are related to the basic information extracted from the text, group D consists of 36 POS tags based on Penn POS [45], group E captures the linguistic traits of the text and group F is related to the readability of the text. Groups C–E are extracted using the Natural Language Toolkit [46] and additional custom functions and formulas written in Python. The features in group F, representing the readability scores, were extracted using functions from the TextStat project [47].

Subsequently, all extracted features were normalised using the Min-Max Normalisation technique. Without normalisation, the scale range of each feature may differ, which would affect the training process. After normalisation, the data were split to be n-fold and then grouped into either the training set or the test set for each fold. However, the datasets used in our study were imbalanced, which can affect prediction [43, 48, 49]. Hence, we investigated the effect of the dynamic sampling process that we proposed in previous research [44]. The sampling process works dynamically based on the current composition of minority and majority class features immediately before the training process begins. This process decreases the majority class in under-sampling, or increases the minority class in over-sampling, to a new ratio (which can be set). The sampling process was applied to the training sets. For simplicity, in this study, we set the ratio to be 1:1 for both dynamic over- and under-sampling—a block diagram representing this sampling process is provided in Fig. 3 [44].

Dynamic random over- and under-sampling have strengths and weaknesses [44]. The strength of over-sampling is that it provides sufficient samples for the minority class critical for the training process. However, it also creates duplicates, which can lead to overfitting. In contrast, under-sampling does not duplicate the samples, but it works by reducing the majority class. Therefore, it may lead to the deletion of important traits of the majority class. The other weakness of under-sampling is that, if the minority class is too small, it will significantly reduce the majority class samples. This is problematic as ML algorithms need a large amount of data for training.

We used four ensembles of ML classifier models as the inner ensembles of EECs for detecting depression from Twitter posts: Bagging Predictors (BP) [50], Random Forest (RF) [51], Adaptive Boosting (AB) [52] and Gradient Boosting (GB). These ensembles are often used in text analysis and have proved to offer excellent performance in previous studies on textual-based sentiment analysis [42], malicious web domain identification [49] and depression detection [13]. The performance of the EECs testing using combinations of the ensemble models above was assessed using four common measurements for prediction or classification; namely, accuracy, precision, recall and F-measure. All ensembles and measurements were built using scikit-learn components [53]. Default parameters were used for all ensemble models to ensure that the results are only affected by implementing our approach and not by modifying classifier parameters.

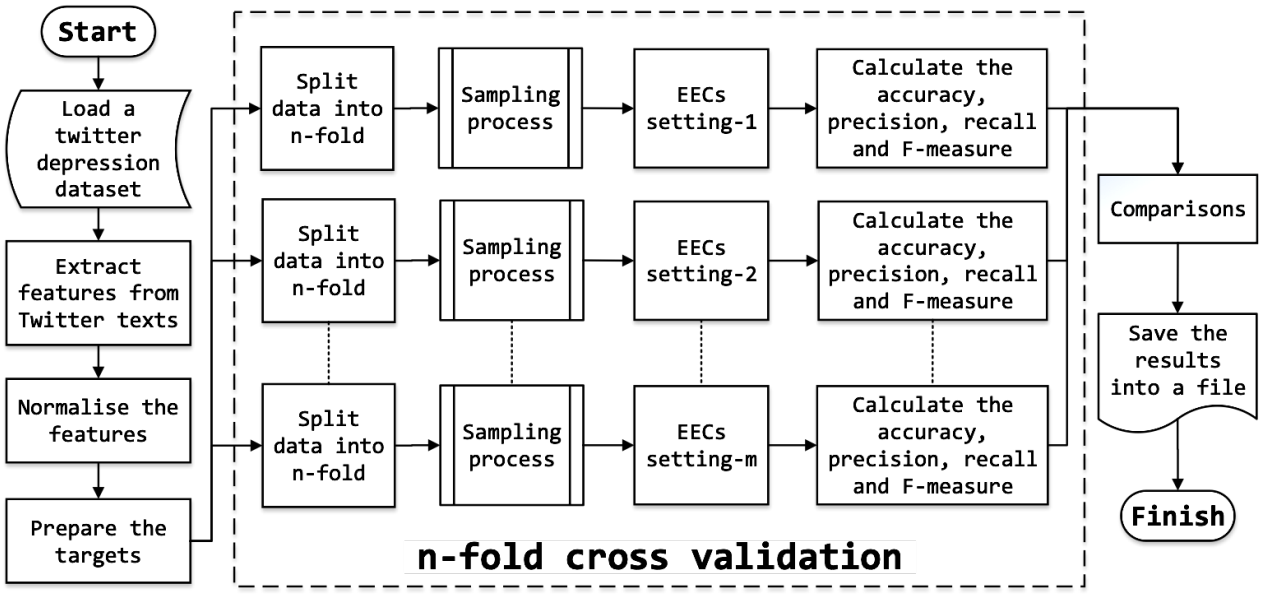

Fig. 2 Framework to test and compare the EECs inner ensemble setting

TABLE 3 FEATURES

No.	Group	Description
Sentiment lexicon features		
1	**A:** Sentiment lexicon features based on SentiWordNet	Total of sentiment terms
2 - 4		Total of (positive, neutral, negative) of sentiment terms
5, 6		The ratio of (positive, negative) sentiment to neutral terms
7		The ratio of negative to positive sentiment terms
8, 9		(Positive, negative) sentiment scores
10	**B:** Sentiment lexicon features based on SenticNet	Total of sentiment terms
11 - 13		Total of (positive, neutral, negative) of sentiment terms
14, 15		The ratio of (positive, negative) sentiment to neutral terms
16		The ratio of negative to positive sentiment terms
17, 18		(Positive, negative) sentiment scores
19		Total introspection value
20		Total temper value
21		Total attitude value
22		Total sensitivity value
Textual Content–based Features		
23 - 26	**C:** Basic text information	Total (letters, words, stop words, sentences) in the review
27		Total words with capitalised 1st letter
28		Total negative terms (e.g., 'does not', 'do not', 'will not')
29		Total elongated words (e.g., 'Yesss', 'fiiine', 'yoouu')
30, 31		Total exclamation and question sentences
32		The existence of weblink inside the text
33 - 68	**D:** POS	Total existence of 36 Tags of Penn POS
69	**E:** Linguistic characteristics	The ratio of adjectives and adverbs
70		Average of number of words per sentence
71		The ratio of word repetition to total words
72		The average number of letters per word
73		Average of words with 1st capital to total sentences.
74		The ratio of words with 1st capital to total words
75 - 77		Total of (1st, 2nd, 3rd) person pronouns
78 - 80		The ratio of (1st, 2nd, 3rd) person pronouns to total pronouns
81 - 87	**F:** Readability scores	Flesch reading ease, Simple Measure of Gobbledygook (SMOG) index, Flesch Kincaid grade, Coleman-Liau index, Gunning fog

		index, Dale–Chall readability and Linsear Write formula
88		Automated readability index (ARI)
89		Difficult words
90		Estimation of school grade level required to understand the text.

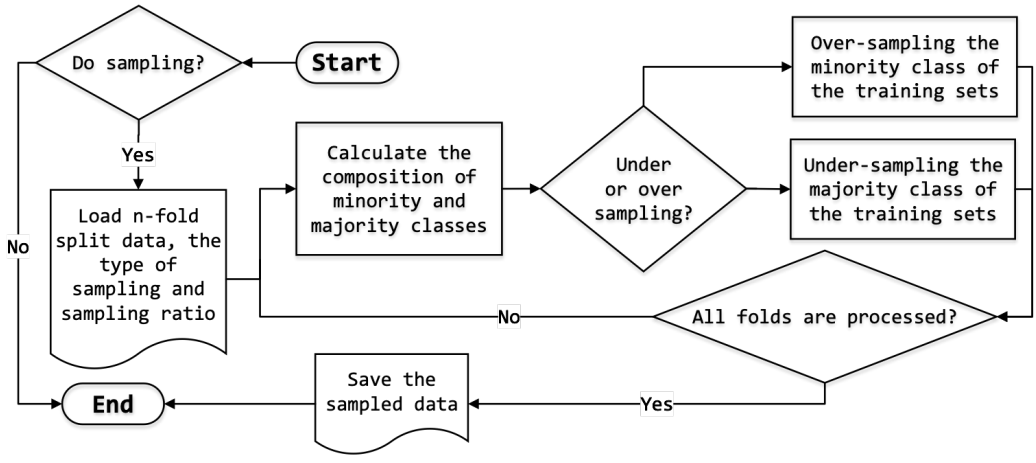

Fig. 3 The sampling process

5 Results and Discussion

5.1 Experimenting on EECs for Twitter's depression detection

The first group of experiments sought to test the performance of EECs regarding depression detection from Twitter messages. For the inner ensembles of EECs, we implemented RF, BP, AB and GB, with parameters set as the scikit-learn component defaults used to implement these classifiers [53]. As shown in Table 4, we tested all possibilities from 2-, 3- to 4-combinations. Besides combining the inner ensembles, other settings were fixed: total of inner ensembles = 25, shared equally for each type; bootstrap = 0.5; and final output = majority vote. The datasets we used in the experiments were drawn from Eye [16] and Shen et al. [15]. All experiments were conducted using the 10-fold CV method. For the comparison, we also presented the results of several well-known ensemble classifiers (RF, BP, AB, GB) on the same dataset as in our previous study [13] (see Table 5). The results were already impressive (all measurements ranged between 94% and 98%); however, we perceived a way of improving these results further using EECs.

As shown in Table 4 (num. 1–4), compared with Table 5, the ensembles, if run inside EECs as inner ensembles, gave slightly better results, except for RF, which is slightly worse. The EECs(AB) performed best for Eye's dataset, while for Shen et al.'s dataset, the best performer was EECs(GB). These results show that combining multiple same-type ensembles in EECs could slightly improve the detection performance, except in the case of RF. However, when combined between different types of inner ensembles, the majority experiment resulted in an improvement, especially for Eye's dataset (see the bolded scores in Table 4); that is, accuracy of EECs(GB-AB-BP-RF) for Eye's = 98.18%. In comparison, in Table 5, GB, AB, BP and RF accuracies were 98.05%, 97.88%, 97.51% and 97.67%, respectively. The best for Shen et al.'s dataset is EECs(GB-AB-BP).

However, not all combinations resulted in the anticipated improvement. A few ensembles degrading the performance of the overall EECs when combined with it. Ultimately, this meant that the performance of the EECs was in the middle of those of its inner ensembles. For example, GB accuracy for Shen et al.'s dataset = 98.05%, while BP = 97.55%, and the combination of both in EECs(GB-BP) = 98%. This implies that BP slightly degraded the performance of GB. In a case like this, it is better to use GB directly. However, in the same case, EECs(GB-BP) for Shen et al.'s dataset improved recall, from 98.59% (GB) and 98.31% (BP) to 98.62%. In this case, using the ECCs is acceptable since recall in binary classification is the same as the accuracy of the positive class (depression class), and thus, higher recall means the method is better able to detect the positive class (depression class).

Overall, the improvement in performances in Table 4 is minor. This is expected since the base performances that we used (from our previous study [13]) are already high (see Table 5). Therefore, slight improvements are acceptable since they move closer to perfection (100%). Also, the fact that improvements occurred in most cases (especially for Eye's dataset) means that such improvements are not coincidental. For the rest of the experiments, we chose one combination from each 1–4 possible combinations: EECs(GB), EECs(AB-BP), EECs(AB-BP-RF) and EECs(GB-AB-BP-RF).

TABLE 4 RESULTS OF COMBINATION OF BASE CLASSIFIER RESULTS OF THE EECS MODEL

Num	Ensemble Setting	Eye's Dataset[1]				Shen et al.'s Dataset[1]			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	EECs(AB)	***98.22***	***95.65***	***96.50***	***96.06***	***97.64***	***97.39***	***98.32***	***97.85***
2	EECs(BP)	***97.75***	***93.35***	***96.75***	***95.02***	***97.66***	***97.29***	***98.46***	***97.87***
3	EECs(GB)	***98.06***	94.82	***96.64***	***95.71***	***98.06***	***97.88***	***98.60***	***98.23***
4	EECs(RF)	97.22	94.58	92.95	93.74	97.71	97.79	98.03	97.91
5	EECs(GB-AB)	***98.14***	***95.35***	***96.36***	***95.84***	98.03	97.84	98.57	98.20
6	EECs(GB-BP)	97.99	94.53	***96.67***	95.58	98.00	97.75	***98.62***	98.18
7	EECs(GB-RF)	98.02	95.04	96.18	95.61	97.97	97.94	98.36	98.15
8	EECs(AB-BP)	***98.05***	94.88	***96.51***	***95.68***	***98.00***	***97.75***	***98.64***	***98.19***
9	EECs(AB-RF)	***97.95***	***95.27***	***95.62***	***95.43***	97.91	97.81	***98.38***	98.09
10	EECs(BP-RF)	***97.96***	94.82	***96.20***	***95.49***	***97.95***	97.85	***98.39***	***98.12***
11	EECs(GB-AB-BP)	***98.13***	95.01	***96.73***	***95.85***	***98.06***	97.80	***98.66***	***98.23***
12	EECs(GB-AB-RF)	***98.14***	***95.26***	***96.48***	***95.85***	97.95	97.94	98.31	98.12
13	EECs(AB-BP-RF)	***98.15***	***95.31***	***96.47***	***95.88***	***97.95***	97.77	***98.50***	***98.13***
14	EECs(GB-AB-BP-RF)	***98.18***	***95.37***	***96.59***	***95.97***	98.00	97.81	98.55	98.18

[1] ***Bold-italic*** = the result of EECs is higher than the result of all its inner ensembles; Normal-black = at least one result of the inner ensemble is higher than the result of EECs.

TABLE 5 RESULTS OF THE ENSEMBLE MODELS USING THE SAME DATASETS FROM THE PREVIOUS STUDY [13]

Num	Classifiers	Eye's Dataset				Shen et al.'s Dataset			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
1	AB	97.88	95.25	95.27	95.26	97.40	97.22	98.05	97.63
2	BP	97.51	93.09	96.01	94.52	97.55	97.23	98.31	97.77
3	GB	98.05	95.11	96.30	95.69	98.05	97.87	98.59	98.22
4	RF	97.67	95.09	94.46	94.77	97.93	98.00	98.22	98.11

5.2 Experimenting with the EECs parameters

In this section, we discuss several input parameters for the EECs that are likely to affect the performance of this ensemble model. We used the same settings as in Section 4.1, except the parameter under investigation. We ran all experiments for the parameter testing on both datasets (see Table 2).

The first investigation explored the effect of the total number of inner ensembles. Here, experiments were run on all four combinations, with the total number of inner ensembles varying from 5 to 65. Results can be seen in Fig. 4 for accuracy, Fig. 5 for precision, Fig. 6 for recall and Fig. 7 for F-measure (please note that we used a different range of X-axis for each chart to better show the detail). From these figures, we can see that accuracy, precision, recall and F-measure was majority increasing. After 25 classifiers, all scores fluctuated around a number, with this fluctuation less than 0.5%. We continued the experiments using the total number of inner ensembles from this set of experiment results (25).

While in Fig. 4, we can see that accuracy for Shen et al.'s and Eye's datasets are similar, for other measures, such as precision in Fig. 5, recall in Fig. 6 and F-measure in Fig. 7, the results for Eye's dataset are lower than for Shen et al.'s dataset. We suspect that Eye's dataset is heavily imbalanced, as depression class records account for only 22% of the total. Therefore, the EECs is more trained to detect the non-depression class than the depression class. This situation is exacerbated because the Shen et al. dataset is slightly imbalanced in favour of the depression class (see Table 2). To be sure that this logic is correct, we also investigated the accuracy of each class. As can be seen in Table 6, the accuracy of the depression class for Eye's dataset is lower than that of the non-depression class. In contrast, for Shen et al.'s dataset, where depression class records are slightly higher than non-depression class records, we can see the accuracy of the depression class is also slightly higher than that of the non-depression class. Thus, we conclude that if we want the positive class (such as the depression class) to be easier to detect than the negative class (non-depression), we need to provide this class with more training samples. However, that is not what happens in reality; in the real world, the positive class usually has fewer samples than the negative class. Later, we discuss how to improve the

performance of the positive class, when the dataset is imbalanced against such class and positive class samples are much lower than for the other class.

The subsequent experiments aimed to investigate the bootstrap setting parameter. We used fixed parameters except for the bootstrap setting, which ranged from 0.25 bootstrapping to 1 (no bootstrapping). As shown in Fig. 8, the best accuracy, precision, recall and F-measure were often achieved when the bootstrap setting was 0.5. Every inner ensemble was trained using 50% of training samples randomly. For the next set of experiments, we set the bootstrap to 0.5. However, as shown in Fig. 8, similarly to the previous set of experiments, precision, recall and F-measure for Eye's dataset were lower than for Shen et al.'s dataset.

The last parameter to test is how the output of the inner ensembles should be processed. There are two options for transferring the output of inner ensembles to the final output: majority vote or priority class threshold. To test the effect of priority, we set the priority to the positive class (depression class), in the range between absolute priority to 45% priority. Absolute priority will give a 'positive' (depression) result if one or more of the inner ensemble suggests; percentage priority will give a 'positive' (depression) result if the number of inner ensembles with positive (depression) results is the same or more than the chosen percentage threshold. The results of these experiments can be seen in Fig. 9.

As per Fig. 9, priority setting increases recall but decreases other measurements. However, the higher the recall means, the more accurate the positive class (depression class) detection is. To investigate further, we also measured the accuracy of each class (see Fig 10). As shown in Fig. 10, prioritising the depression class (positive class) when processing the output of inner ensembles increases the accuracy of this class and decreases the accuracy of the non-depression class (negative class). Therefore, the key here is how we choose the percentage of priority so that the accuracy increase of the positive class does not greatly decrease that of the negative class. For example, if we set 25% priority, the accuracy of the depression (positive) class increases in almost all cases, but the accuracy of the non-depression (negative) class did not decrease much compared with 15%, 5% and absolute priority.

TABLE 6 COMBINED AND DETAILED ACCURACY OF EECS COMBINATIONS

Dataset	**EECs combination**	**Combined Accuracy**	**Detailed accuracy of each class**	
			Depression	**Non-depression**
Eye's	EECs(GB)	98.06	96.64	98.47
	EECs(AB-BP)	98.05	96.51	98.50
	EECs(AB-BP-RF)	98.15	96.47	98.64
	EECs(GB-AB-BP-RF)	98.18	96.59	98.64
Shen et al.'s	EECs(GB)	98.06	98.60	97.42
	EECs(AB-BP)	98.00	98.64	97.24
	EECs(AB-BP-RF)	97.95	98.50	97.29
	EECs(GB-AB-BP-RF)	98.00	98.55	97.35

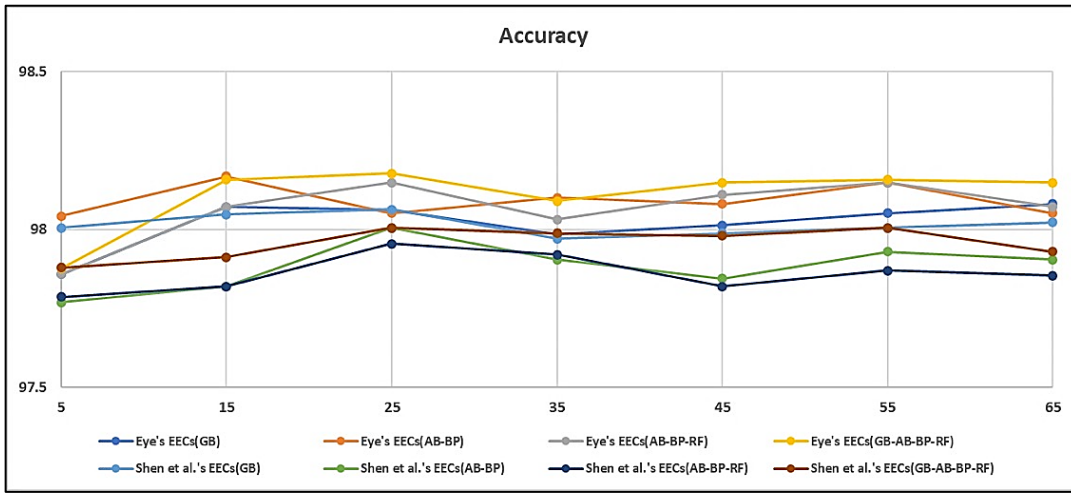

Fig. 6 The accuracy of the EECs combinations with 5–65 total inner ensembles (x-axis is accuracy in percentage; y-axis is total inner ensembles)

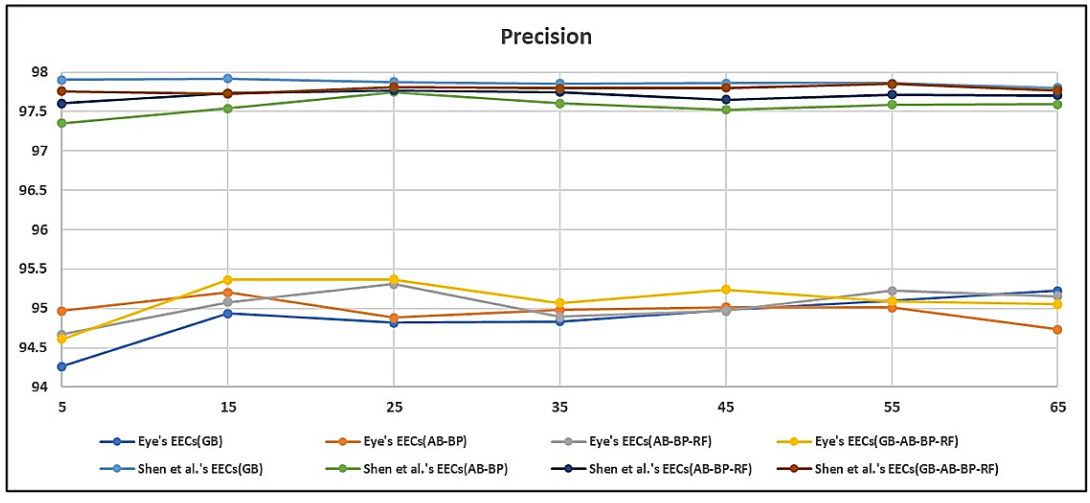

Fig. 5 The precision of the EECs combinations with 5–65 total inner ensembles (x-axis is precision in percentage; y-axis is total inner ensembles)

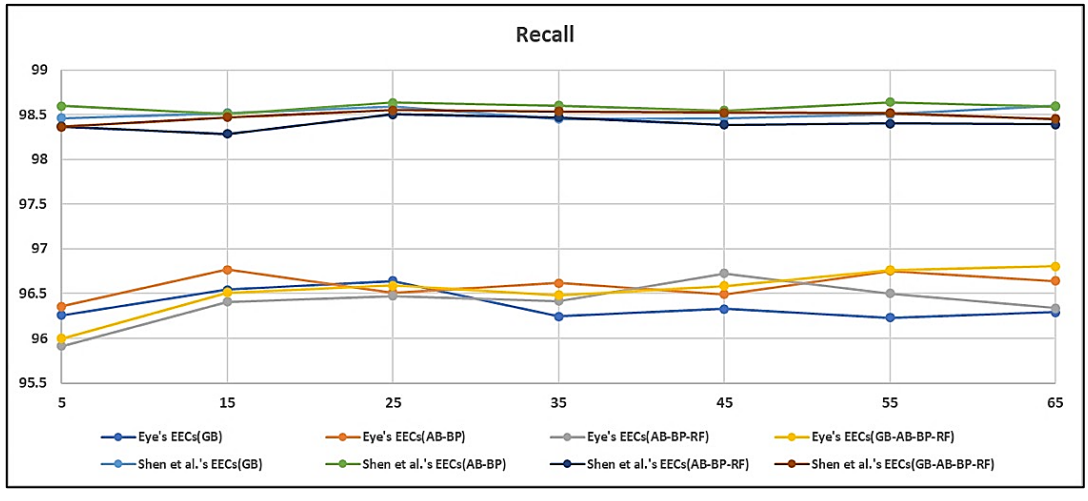

Fig. 4 The recall of the EECs combinations with 5–65 total inner ensembles (x-axis is recall in percentage; y-axis is total inner ensembles)

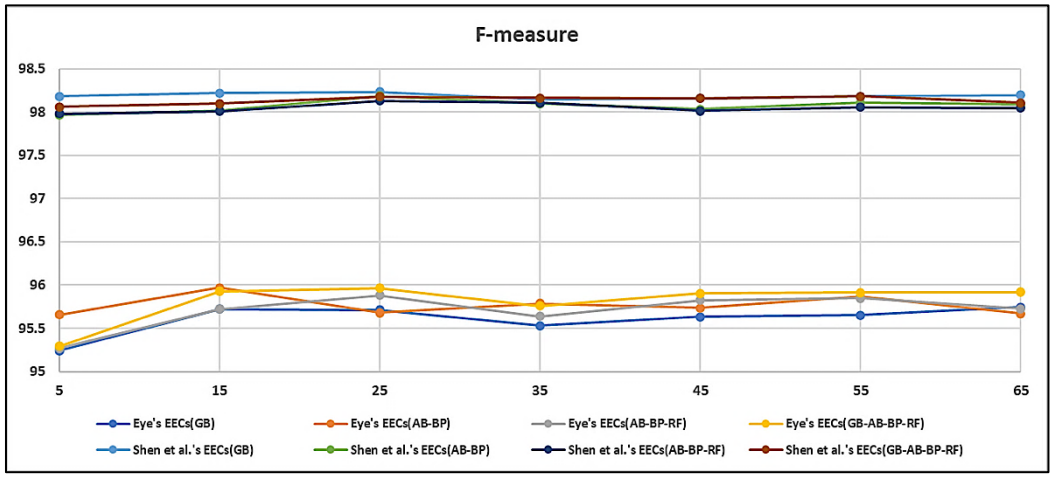

Fig. 7 The F-measure of the EECs combinations with 5–65 total inner ensembles (x-axis is F-measure in percentage; y-axis is total inner ensembles)

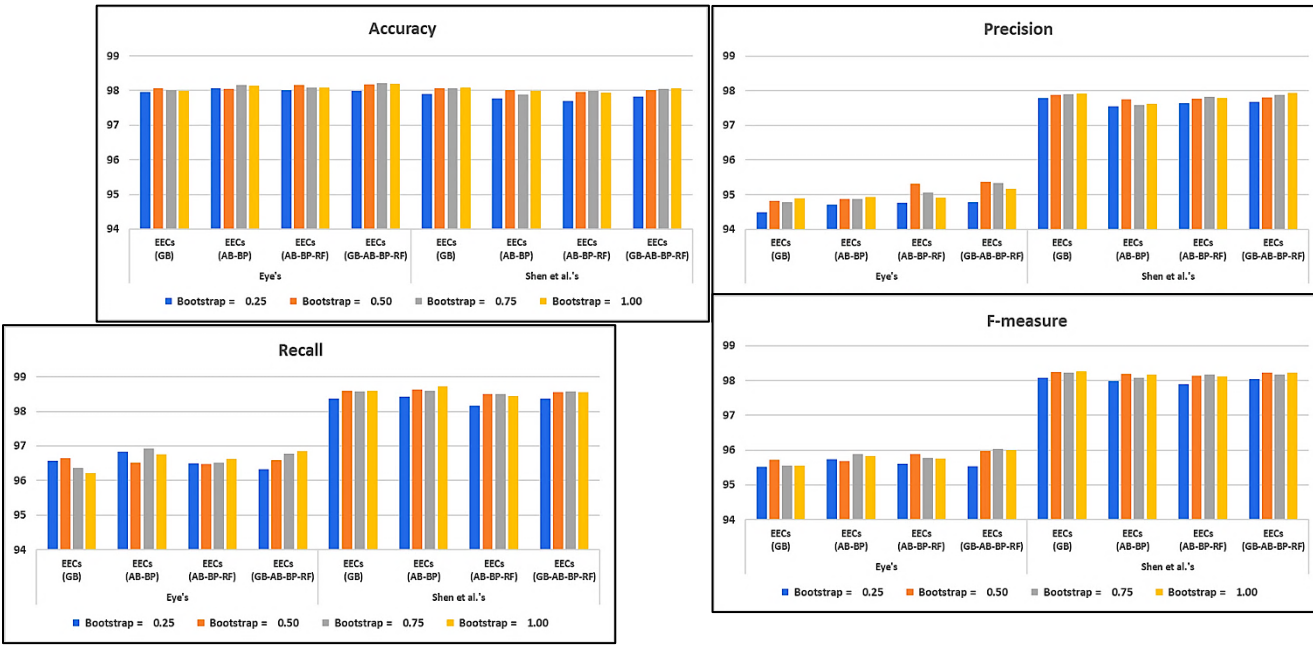

Fig. 8 Accuracy, precision, recall and F-measure of the EECs with bootstrap setting from 0.25 to 1.00 (x-axis = accuracy, precision, recall or F-measure in percentage; y-axis = several settings of EECs inner ensembles of Eye's and Shen et al.'s datasets)

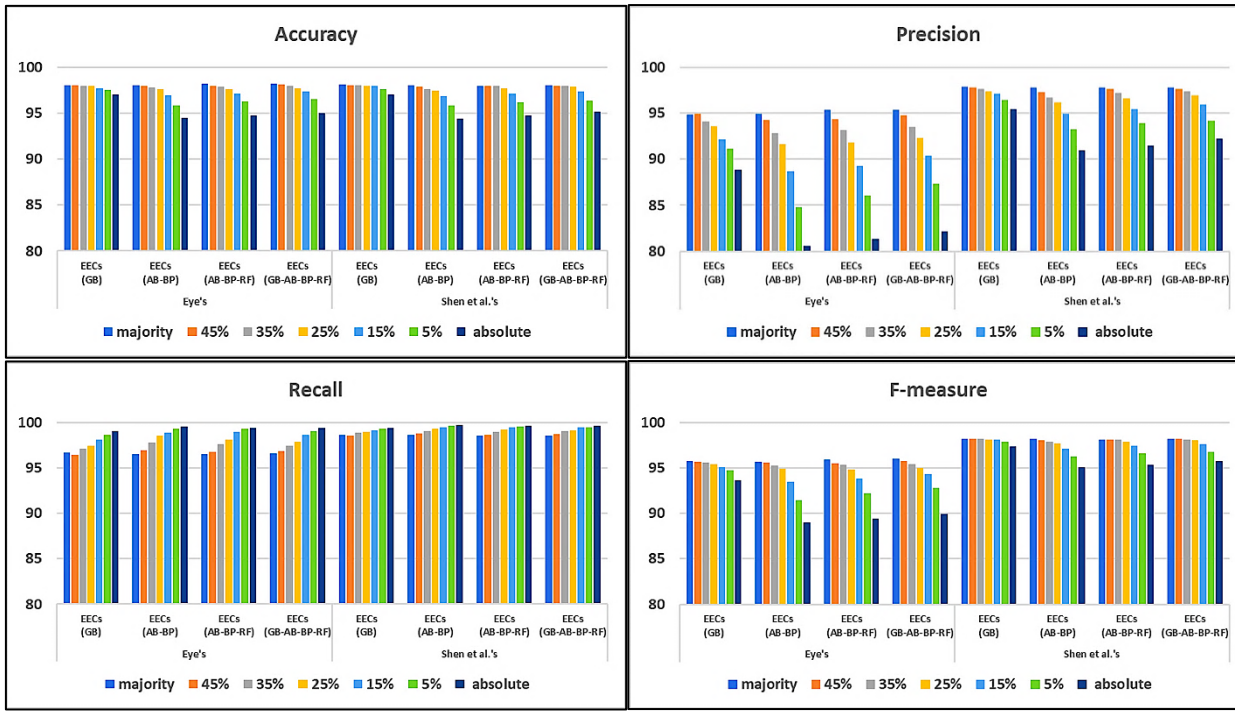

Fig. 9 Results of final target experiments, from majority vote, 45% priority to absolute priority (x-axis = accuracy, precision, recall or F-measure in percentage; y-axis = several settings of EECs inner ensembles of Eye's and Shen et al.'s datasets)

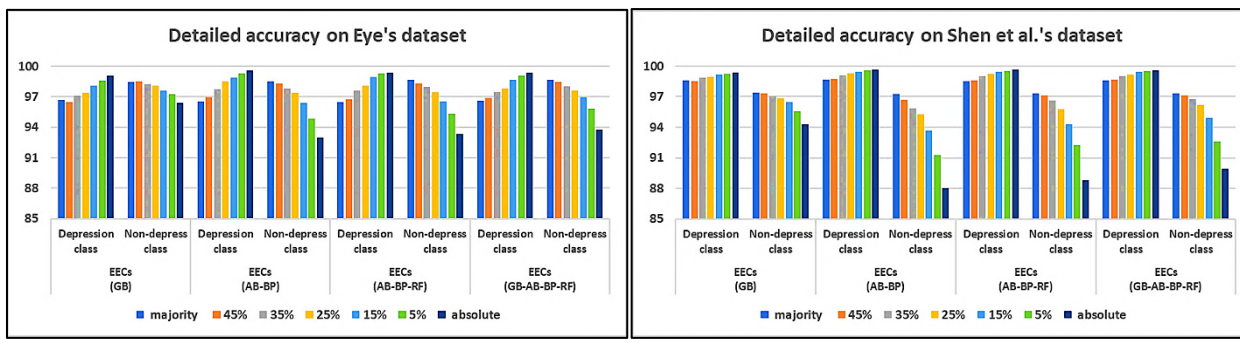

Fig. 10 Detailed accuracy of each class in each dataset, from majority vote, 45% priority to absolute priority (x-axis = detailed accuracy in percentage; y-axis = several settings of EECs inner ensembles of Eye's and Shen et al.'s datasets)

5.3 Effect of dynamic sampling

As noted previously, imbalanced data may affect prediction performances (i.e. accuracy, precision or recall). Therefore, in this subsection, we discuss the effect of dynamic sampling, which we proposed and applied with success in previous research [44]. As shown in Table 2, Eye's dataset is heavily imbalanced in that total depression class records are only 22% of all records. In contrast, the depression class records are slightly higher in Shen et al.'s dataset, accounting for 54% of total records.

As shown in Fig. 11, both over- and under-sampling of the training data have a significant effect for Eye's dataset, which is heavily imbalanced. The accuracy of the depression class (minority) increases 2% to 2.5%, from 96% to almost 99%, in all EECs combinations. However, the accuracy of the non-depression class (the majority class) is reduced. In all cases, under-sampling gives a higher increase for the depression class but decreases the non-depression class. In our opinion, this is acceptable for depression detection, for which it is better that the model is more accurate in detecting depression (its purpose) even if less accurate in detecting non-depression. For Shen et al.'s dataset, which is only slightly imbalanced for the non-depression class, the sampling process did not have a significant effect.

For the 25% priority threshold for output setting, as shown in Fig. 12, the sampling process still produces a good effect, as the improvement in depression class accuracy is around 0.5%–2%. In the cases where there is only 0.5%–1% improvement (EECs(AB-BP) and EECs(AB-BP-RF)), this arises because the baselines are already high (more than 98%), implying it would be impossible to attain

an improvement of 2%. Again, the sampling did not have a significant effect for Shen et al.'s dataset, providing proof that the dynamic sampling process can only produce a good result for a heavily imbalanced dataset.

For the absolute priority setting (see Fig. 13), the sampling process only slightly increases the accuracy of Eye's depression class to a maximum of 0.7% in EECs (AB-BP-RF). The baseline of depression class accuracy with the absolute priority is already too high (all above 99%), and this 0.7% increase should be considered a very good result. However, these slight increases in depression class accuracy are accompanied by a significant reduction in non-depression class accuracy. For Shen et al.'s dataset, dynamic sampling has only a trivial effect.

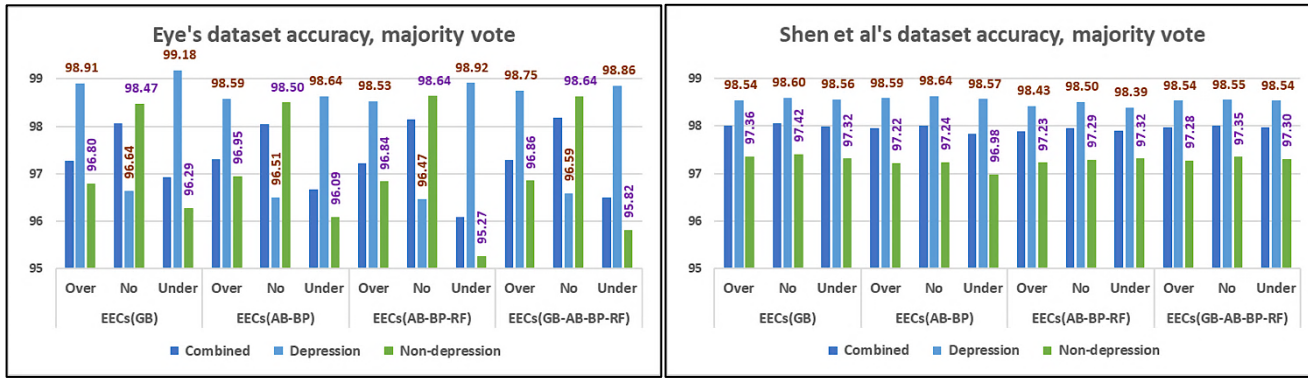

*Fig. 11 The accuracy results of over- and under-sampling for each dataset with EECs output setting = **majority vote** (Over = Over-sampling; No = No-sampling; Under = Under-sampling; x-axis = accuracy in percentage; y-axis = several settings of EECs inner ensembles)*

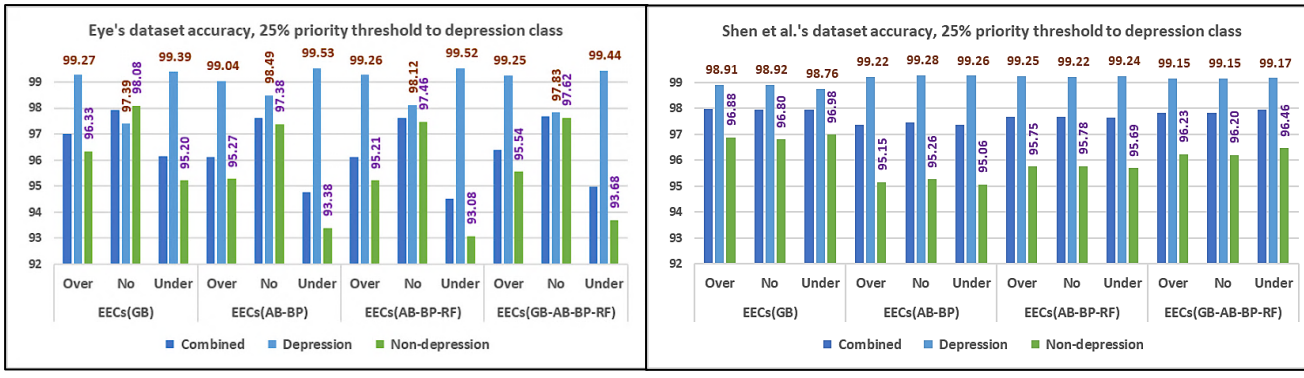

*Fig. 13 The accuracy results of over- and under-sampling for each dataset with EECs output setting = **25% priority** on depression class (Over = Over-sampling; No = No-sampling; Under = Under-sampling; x-axis = accuracy in percentage; y-axis = several settings of EECs inner ensembles)*

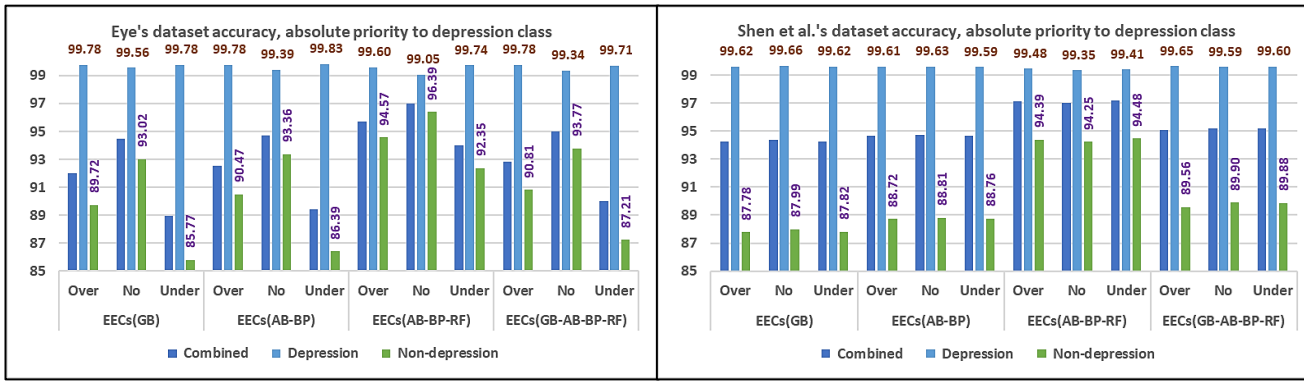

*Fig. 12 The accuracy results of over- and under-sampling for each dataset with EECs output setting = **absolute priority** on depression class (Over = Over-sampling; No = No-sampling; Under = Under-sampling; x-axis = accuracy in percentage; y-axis = several settings of EECs inner ensembles)*

5.4 Comparison to previous work

We cannot compare our results to other studies in Table 1 since the datasets used are different, and the formulas for measurements may also differ. However, we can confidently compare our current results to our previous study [13, 14] (see Table 7). In addition to the best results from our previous study, in Table 7, we also present the best results of our current study for different output settings (majority vote or priority class threshold), combined with the sampling process. For Shen et al.'s dataset, we do not include a combination of majority vote and sampling because, as discussed above, the results are similar. The sampling process did not have a substantially different effect for this slightly imbalanced dataset. For the baseline, we use the best measurements from our previous study [13] (underlined); whenever our current results are better than the baseline, we apply **bold** font. As can be seen in Table 7, the recall results are always the same as the detailed accuracy for the depression class. We calculate these using different components of scikit-learn [53]. For the binary classification, recall produces the same score as positive class accuracy (in this case, depression class). For the combined accuracy, precision and F1, we also calculate these using scikit-learn components.

As can be seen in Table 7 (desc. a), EECs using the majority vote setting, without sampling, only increases the measurements slightly because the baselines are already high. However, the implementation of the priority threshold setting (see Table 7, desc. b and c) has a good impact on recall (or the detailed accuracy of the depression class), which increases 3% for Eye's dataset and 1% for Shen et al.'s dataset. For both datasets, the absolute priority to the depression class can increase the accuracy of depression class detection to more than 99%.

As discussed above, the sampling process did not affect the Shen et al. dataset, which is only slightly imbalanced. Therefore, in Table 7, we only present the best sampling results for Eye's dataset, which is heavily imbalanced. As we can see in Table 7, 1.d–1.f compared with 1.a–1.c, the sampling process improved the recall, or the accuracy of depression class (as a minority), to a maximum of 2.1% for the majority vote and to a minimum of 0.5% for the absolute priority setting. Therefore, the combination of priority setting and the dynamic sampling process improved recall (depression class accuracy) to 99.83% in Eye's dataset and 99.66% in Shen et al.'s dataset, which is almost the maximum possible. In our opinion, this is crucial since the focus of this study is to detect signs of depression in social media.

TABLE 7 BEST RESULTS OF THE EECS AND OUR PREVIOUS STUDIES FOR THE SAME TWITTER DEPRESSION DATASETS

No	Dataset	Description	Com. Acc* (%)	Pre* (%)	Rec* (%)	F1* (%)	Detailed accuracy*	
							D	ND
1	Eye's dataset [16]	Chiong et al. [14], textual-based features, LR	92.61	93.32	72.21	81.38	72.21	98.51
		Chiong et al. [13], hybrid sentiment- and textual content–based features, GB	98.05	95.11	96.30	95.69	96.30	98.57
		*a. EECs(AB), 25, 0.5, Majority-vote, No-sampling***	**98.22**	**95.65**	**96.50**	**96.06**	**96.50**	**98.72**
		*b. EECs(AB-BP), 25, 0.5, 25%-priority, No-sampling***	*97.62*	*91.58*	**98.49**	*94.89*	**98.49**	*97.38*
		*c. EECs(GB-AB-BP-RF), 25, 0.5, Absolute-priority, No-sampling***	*95.02*	*82.17*	**99.34**	*89.91*	**99.34**	*93.77*
		*d. EECs(AB-BP), 25, 0.5, Majority-vote, Over-sampling***	*97.31*	*90.30*	**98.59**	*94.25*	**98.59**	*96.95*
		*e. EECs(GB), 25, 0.5, 25%-priority, Over-sampling***	*96.99*	*88.71*	**99.27**	*93.69*	**99.27**	*96.33*
		*f. EECs(AB-BP), 25, 0.5, Absolute-priority, Under-sampling***	*89.39*	*67.94*	**99.83**	*80.80*	**99.83**	*86.39*
2	Shen et al.'s dataset [15]	Chiong et al. [14], textual-based features, LR	88.62	92.63	86.03	89.2	86.03	91.75
		Chiong et al. [13], hybrid sentiment- and textual content–based features, GB	98.05	97.87	98.59	98.22	98.59	97.41
		*a. EECs(GB), 25, 0.5, Majority-vote, No-sampling***	**98.06**	**97.88**	**98.60**	**98.23**	**98.60**	**97.42**
		*b. EECs(AB-BP), 25, 0.5, 25%-priority, No-sampling***	*97.46*	*96.20*	**99.28**	*97.71*	**99.28**	*95.26*
		*c. EECs(GB), 25, 0.5, Absolute-priority, No-sampling***	*94.37*	*90.91*	**99.66**	*95.08*	**99.66**	*87.99*

** Com. Acc = Combined Accuracy, Pre = precision, Rec = recall, and F1 = F-measure; D = Depression class, ND = Non-depression class*
*** EECs(<type of inner ensembles>, <total inner ensembles>, <bootstrap percentage>, <majority or priority output>, <sampling type>)*

6 Conclusion

Mental disorders, especially depression, can be dangerous if left untreated. Depression reduces quality of life and, even worse, is the main cause of suicide. Unfortunately, many cases go untreated because of a failure to detect as well as a denial of the condition. However, studies have shown that, with the exponential increase in social media usage, social media messages could provide a valuable source for monitoring mental health issues, including depression.

From the results, we conclude that our proposed ensemble, the EECs, can accurately detect depression using Twitter text. Compared with our previous approach, the EECs can improve recall or accuracy of the positive class to more than 99.5%. Not all combinations of inner ensembles produced the expected result of an improvement in the performance. However, when the combinations were correct, the EECs could produce better results than when the inner ensembles are run alone.

The number of inner ensembles affected performance detection. While reducing the amount of data for the training process, the bootstrap setting also affected the performance of the EECs. Analysing the experiments, we found that 50% bootstrapping gives better results than other percentages. The majority vote setting for the output offers a fair chance for each class to be detected; however, the priority threshold boosts detection of the prioritised class. This would be useful for depression detection since it boosts the accuracy of the priority class (depression) to more than 99.5%, which is almost perfect. However, its weakness is that it also reduces the accuracy of the unprioritised class. The sampling process affected the heavily imbalanced dataset by improving the minority class detection accuracy; however, little impact was recorded for the only slightly imbalanced dataset.

Despite the extensive experimental studies presented in this paper, there remain opportunities/possibilities for further improvement. For our future work, we plan to investigate the implementation of the EECs for other mental disorders and other types of datasets. We hope that our proposed depression detection model can be used by psychologists and psychiatrists via a mobile application to help monitor their patients' conditions, make diagnoses, predict treatment outcomes and select the correct treatments.

Acknowledgment

This study was funded by the University of Newcastle's College of Engineering, Science and Environment's Multidisciplinary Strategic Investment Scheme.

Corresponding authors: Gregorius Satia Budhi, Raymond Chiong

Raymond Chiong graduated with a PhD degree from the University of Melbourne, Australia. He is currently an associate professor with the School of Information and Physical Sciences, University of Newcastle, Australia. He has been actively pursuing research related to the use of automated intelligent computing methods, including machine learning and optimisation algorithms. To date, He has produced over 220 refereed publications and attracted more than 3 million dollars in research and industry funding. He is the Editor-in-Chief of the Journal of Systems and Information Technology and an Editor of Engineering Applications of Artificial Intelligence.

Gregorius Satia Budhi graduated with a PhD degree from the University of Newcastle, Australia. He is currently an associate professor with Petra Christian University in Indonesia. His research interests include sentiment analysis, machine learning and data/text mining. He has more than 70 publications to date in these areas.

Erik Cambria is a professor of computer science and engineering at Nanyang Technological University, Singapore. He obtained his PhD through a joint programme between the University of Stirling and MIT Media Lab. His research focuses on neurosymbolic AI for explainable natural language processing in domains like sentiment analysis, dialogue systems, and financial forecasting. He is an IEEE Fellow, and Associate Editor of many top-tier AI journals.

References

[1] R. A. Adams, Q. J. Huys, and J. P. Roiser, "Computational Psychiatry: Towards a mathematically informed understanding of mental illness," *Journal of Neurology, Neurosurgery and Psychiatry,* vol. 87, no. 1, pp. 53-63, Jan 2016, doi: https://doi.org/10.1136/jnnp-2015-310737.

[2] H. S. Alsagri and M. Ykhlef, "Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features," *IEICE Transactions on Information and Systems,* vol. E103.D, no. 8, pp. 1825-1832, 2020, doi: https://doi.org/10.1587/transinf.2020EDP7023.

[3] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications, vol. 3*4, pp. 10309–10319, 2022, doi: https://doi.org/10.1007/s00521-021-06208-y.

[4] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems,* vol. 8, no. 1, pp. 214-226, 2021, doi: https://doi.org/10.1109/TCSS.2020.3021467.

[5] S. Nikolin, Y. Y. Tan, A. Schwaab, A. Moffa, C. K. Loo, and D. Martin, "An investigation of working memory deficits in depression using the n-back task: A systematic review and meta-analysis," *Journal of Affective Disorders,* vol. 284, pp. 1-8, 2021/04/01/ 2021, doi: https://doi.org/10.1016/j.jad.2021.01.084.

[6] S. Yadav, T. Kaim, S. Gupta, U. Bharti, and P. Priyadarshi, "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: https://doi.org/10.1109/ICACCCN51052.2020.9362738.

[7] N. S. Srimadhur and S. Lalitha, "An End-to-End Model for Detection and Assessment of Depression Levels using Speech," *Procedia Computer Science,* vol. 171, pp. 12-21, 2020, doi: https://doi.org/10.1016/j.procs.2020.04.003.

[8] P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan, "Computational psychiatry," *Trends in Cognitive Sciences,* vol. 16, no. 1, pp. 72-80, Jan 2012, doi: https://doi.org/10.1016/j.tics.2011.11.018.

[9] Q. J. Huys, T. V. Maia, and M. J. Frank, "Computational psychiatry as a bridge from neuroscience to clinical applications," *Nature Neuroscience,* vol. 19, no. 3, pp. 404-13, Mar 2016, doi: https://doi.org/10.1038/nn.4238.

[10] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq, and S. Lee, "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: https://doi.org/10.1109/ICTC.2017.8190959.

[11] J. Hussain *et al.*, "Exploring the dominant features of social media for depression detection," *Journal of Information Science,* vol. 46, no. 6, pp. 739-759, 2019, doi: https://doi.org/10.1177/0165551519860469.

[12] S. Han, R. Mao, and E. Cambria, "Hierarchical Attention Network for Explainable Depression Detection on Twitter aided by Metaphor Concept Mappings," in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, October 12–17, 2022, pp. 94-104. [Online]. Available: https://aclanthology.org/2022.coling-1.9.

[13] R. Chiong, G. S. Budhi, and S. Dhakal, "Combining sentiment lexicons and content-based features for depression detection," *IEEE Intelligent Systems,* vol. 36, no. 6, 2021, doi: https://doi.org/10.1109/MIS.2021.3093660.

[14] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Computers in Biology and Medicine,* vol. 135, 104499, 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104499.

[15] G. Shen *et al.*, "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.

[16] B. B. Eye. *Depression Analysis*. [Online]. Available: https://www.kaggle.com/bababullseye/depression-analysis

[17] A. Kato, Y. Kunisato, K. Katahira, T. Okimura, and Y. Yamashita, "Computational Psychiatry Research Map (CPSYMAP): A new database for visualising research papers," *Frontiers in Psychiatry,* vol. 11, p. 578706, 2020, doi: https://doi.org/10.3389/fpsyt.2020.578706.

[18] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Procedia Computer Science,* vol. 167, pp. 1258-1267, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.442.

[19] P. Kumar, S. Garg, and A. Garg, "Assessment of anxiety, depression and stress using machine learning models," *Procedia Computer Science,* vol. 171, pp. 1989-1998, 2020, doi: https://doi.org/10.1016/j.procs.2020.04.213.

[20] B. Ojeme and A. Mbogho, "Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders," *Procedia Computer Science,* vol. 96, pp. 1294-1303, 2016, doi: https://doi.org/10.1016/j.procs.2016.08.174.

[21] T. V. Wiecki, J. Poland, and M. J. Frank, "Model-Based Cognitive Neuroscience Approaches to Computational Psychiatry," *Clinical Psychological Science,* vol. 3, no. 3, pp. 378-399, 2015, doi: https://doi.org/10.1177/2167702614565359.

[22] J. R. Sato, J. Moll, S. Green, J. F. Deakin, C. E. Thomaz, and R. Zahn, "Machine learning algorithm accurately detects fMRI signature of vulnerability to major depression," *Psychiatry Res,* vol. 233, no. 2, pp. 289-91, Aug 30 2015, doi: https://doi.org/10.1016/j.pscychresns.2015.07.001.

[23] K. H. Brodersen *et al.*, "Dissecting psychiatric spectrum disorders by generative embedding," *Neuroimage Clin,* vol. 4, pp. 98-111, 2014, doi: https://doi.org/10.1016/j.nicl.2013.11.002.

[24] J. Gratch *et al.*, "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. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf.

[25] A. Samareh, Y. Jin, Z. Wang, X. Chang, and S. Huang, "Detect depression from communication: How computer vision, signal processing, and sentiment analysis join forces," *IISE Transactions on Healthcare Systems Engineering,* vol. 8, no. 3, pp. 196-208, 2018, doi: https://doi.org/10.1080/24725579.2018.1496494.

[26] Q. Chen, I. Chaturvedi, S. Ji, and E. Cambria, "Sequential fusion of facial appearance and dynamics for depression recognition," *Pattern Recognition Letters,* vol. 150, pp. 115-121, 2021, doi: https://doi.org/10.1016/j.patrec.2021.07.005.

[27] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific reports,* vol. 10, no. 1, pp. 11846-11846, 2020, doi: https://doi.org/10.1038/s41598-020-68764-y.

[28] Y. Chen, B. Zhou, W. Zhang, W. Gong, and G. Sun, "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: https://doi.org/10.1109/dsc.2018.00073.

[29] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health Inf Sci Syst,* vol. 6, no. 1, p. 8, Dec 2018, doi: https://doi.org/10.1007/s13755-018-0046-0.

[30] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Expert Systems with Applications,* vol. 133, pp. 182-197, 2019, doi: https://doi.org/10.1016/j.eswa.2019.05.023.

[31] I. Fatima, B. U. D. Abbasi, S. Khan, M. Al-Saeed, H. F. Ahmad, and R. Mumtaz, "Prediction of postpartum depression using machine learning techniques from social media text," *Expert Systems,* vol. 36, no. 4, 2019, doi: https://doi.org/10.1111/exsy.12409.

[32] E. M. S. Filho *et al.*, "Can machine learning be useful as a screening tool for depression in primary care?," *J Psychiatr Res,* vol. 132, pp. 1-6, Jan 2021, doi: https://doi.org/10.1016/j.jpsychires.2020.09.025.

[33] N. Jothi, W. Husain, and N. A. Rashid, "Predicting generalised anxiety disorder among women using Shapley value," *J Infect Public Health,* vol. 14, no. 1, pp. 103-108, Apr 6 2020, doi: https://doi.org/10.1016/j.jiph.2020.02.042.

[34] H. Jung, H. A. Park, and T. M. Song, "Ontology-Based Approach to Social Data Sentiment Analysis: Detection of Adolescent Depression Signals," *J Med Internet Res,* vol. 19, no. 7, p. e259, Jul 24 2017, doi: https://doi.org/10.2196/jmir.7452.

[35] B. Sutter, R. Chiong, G. S. Budhi, and S. Dhakal, "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, doi: https://doi.org/10.1007/978-3-030-79457-6_30.

[36] C. Lin *et al.*, "SenseMood: Depression Detection on Social Media," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020: Association for Computing Machinery, pp. 407–411, doi: https://doi.org/10.1145/3372278.3391932.

[37] S. Baccianella, A. Esuli, and F. Sebastian, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta, May 17-23, 2010, vol. 10, pp. 2200-2204. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf.

[38] E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok, "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," in *Proceedings of the International Conference on Human-computer Interaction (HCII)*, Washington DC, USA, 29 June - 4 July, 2024.

[39] S. Virahonda. *Depression and anxiety comments*, Kaggle. [Online]. Available: https://www.kaggle.com/sergiovirahonda/depression-and-anxiety-comments

[40] R. Tanwar. *Victoria Suicide Data*, Kaggle. [Online]. Available: https://www.kaggle.com/ravijoe/victoria-suicide-data

[41] M. Valstar *et al.*, "Avec 2014 - 3D dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge - AVEC '14*, Orlando, Florida, USA, November 7, 2014, pp. 3-10, doi: https://doi.org/10.1145/2661806.2661807.

[42] G. S. Budhi, R. Chiong, I. Pranata, and Z. Hu, "Using machine learning to predict the sentiment of online reviews: A new framework for comparative analysis," *Archives of Computational Methods in Engineering,* vol. 28, pp. 2543–2566, 2021, doi: https://doi.org/10.1007/s11831-020-09464-8.

[43] G. S. Budhi, R. Chiong, and Z. Wang, "Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features," *Multimedia Tools and Applications,* vol. 80, pp. 13079-13097, 2021, doi: https://doi.org/10.1007/s11042-020-10299-5.

[44] G. S. Budhi, R. Chiong, Z. Wang, and S. Dhakal, "Using a hybrid content-based and behaviour-based featuring approach in a parallel environment to detect fake reviews," *Electronic Commerce Research and Applications,* vol. 47, 101048, 2021, doi: https://doi.org/10.1016/j.elerap.2021.101048.

[45] S. Buchholz, "Memory-based grammatical relation finding," Tilburg, Eigen beheer, 2002. [Online]. Available: https://pure.uvt.nl/ws/portalfiles/portal/1061771/3238957.pdf

[46] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. USA: O'Reilly Media, Inc., 2009.

[47] S. Bansal and C. Aggarwal. "textstat 0.5.6." https://pypi.org/project/textstat/#description (accessed October 2, 2019).

[48] D. Zhang, L. Zhou, J. L. Kehoe, and I. Y. Kilic, "What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews," *Journal of Management Information Systems,* vol. 33, no. 2, pp. 456-481, 2016, doi: https://doi.org/10.1080/07421222.2016.1205907.

[49] Z. Hu, R. Chiong, I. Pranata, Y. Bao, and Y. Lin, "Malicious web domain identification using online credibility and performance data by considering the class imbalance issue," *Industrial Management & Data Systems,* vol. 119, no. 3, pp. 676-696, 2019, doi: https://doi.org/10.1108/IMDS-02-2018-0072.

[50] L. Breiman, "Bagging predictors," *Machine Learning,* vol. 24, no. 2, pp. 123-140, 1996, doi: https://doi.org/10.1007/bf00058655.

[51] L. Breiman, "Random forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001, doi: https://doi.org/10.1023/a:1010933404324.

[52] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences,* vol. 55, no. 1, pp. 119-139, 1997/08/01/ 1997, doi: https://doi.org/10.1006/jcss.1997.1504.

[53] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825-2830, May 08 2011. [Online]. Available: http://scikit-learn.org/stable/modules/classes.html.