Int. J. Appl. Math. Comput. Sci., 2023, Vol. 33, No. 1, 117–131
DOI: 10.34768/amcs-2023-0010

A CONTEMPORARY MULTI–OBJECTIVE FEATURE SELECTION MODEL FOR DEPRESSION DETECTION USING A HYBRID pBGSK OPTIMIZATION ALGORITHM

SANTHOSAM KAVI PRIYA [a,*], KASIRAJAN PON KARTHIKA [a]

[a] Department of Computer Science and Engineering
Mepco Schlenk Engineering College (Autonomous)
Sivakasi, 626005, Tamil Nadu, India
e-mail: urskavi@mepcoeng.ac.in

Depression is one of the primary causes of global mental illnesses and an underlying reason for suicide. The user generated text content available in social media forums offers an opportunity to build automatic and reliable depression detection models. The core objective of this work is to select an optimal set of features that may help in classifying depressive contents posted on social media. To this end, a novel multi-objective feature selection technique (EFS-pBGSK) and machine learning algorithms are employed to train the proposed model. The novel feature selection technique incorporates a binary gaining-sharing knowledge-based optimization algorithm with population reduction (pBGSK) to obtain the optimized features from the original feature space. The extensive feature selector (EFS) is used to filter out the excessive features based on their ranking. Two text depression datasets collected from Twitter and Reddit forums are used for the evaluation of the proposed feature selection model. The experimentation is carried out using naive Bayes (NB) and support vector machine (SVM) classifiers for five different feature subset sizes (10, 50, 100, 300 and 500). The experimental outcome indicates that the proposed model can achieve superior performance scores. The top results are obtained using the SVM classifier for the SDD dataset with 0.962 accuracy, 0.929 F1 score, 0.0809 log-loss and 0.0717 mean absolute error (MAE). As a result, the optimal combination of features selected by the proposed hybrid model significantly improves the performance of the depression detection system.

Keywords: depression detection, text classification, dimensionality reduction, hybrid feature selection, binary gaining-sharing knowledge-based optimization.

1. Introduction

The global pandemic has brought about a wide range of serious physical, emotional and economic problems, leading to a variety of mental health issues such as anxiety, post traumatic stress disorder, etc. Among them, depression is one of the serious mental health issues that needs attention since it may threaten human lives. Depression disorder (Friedrich, 2017) has a strong effect on all sorts of individuals in all aspect of life including productivity and performance at work, personal and family relationships, friendships and social communications. Treatment and support services are generally absent in underdeveloped countries. About 76–85% of individuals affected by mental disorders in these countries remain untreated due to the lack of access to the treatment they need. It can be easily cured when it is detected at the early stage and a proper diagnosis or guidance has to be provided.

Finding depression at an early stage is a challenging task. Conventionally, depression is detected and diagnosed using self-report questionnaires by practising clinicians. The depressed individuals most often hesitate to seek medical assistance due to the social slur around the myths of depression. The advanced Internet communication technologies enabled the depressed individuals to share their thoughts, challenges and experiences about the mental health disorders via social media, blogs or online forums. The content expressed by the individual in various social media forums provides an opportunity to develop depression detection systems using state-of-the-art natural language processing (NLP)

*Corresponding author

and machine learning (ML) techniques. Due to the unstructured nature of the social media texts, ML models are well suited to handle the nonlinearities better than the conventional statistical methods. The proposed work formulates the depression detection problem as a text classification problem using social media posts.

The core contributions of the proposed work are as follows.

1. Developing a new multi-objective hybrid feature selection strategy to classify depression related texts combining a filter method (EFS) and a wrapper method (pBGSK).

2. Employing a filter feature selection approach based on an EFS measure to remove the redundant and irrelevant features and rank the features based on their significance.

3. Adapting the pBGSK algorithm to find an optimal set of features which helps in reducing the dimensionality of the text vectors for the classification process.

4. Performing extensive comparisons of depression classification models to prove the effectiveness of the proposed work.

The rest of the paper is organized as follows. Section 2 presents recent works related to feature selection methods and depression detection techniques using social media content. Section 3 provides basic concepts of text feature selection, the EFS filter approach and the GSK algorithm. This section also describes the proposed feature selection algorithm based on the EFS-pBGSK. Section 4 explains the experimental settings including datasets, dataset preprocessing, classifiers and evaluation metrics used to measure the performance of the proposed work. The experimental results and a discussion to prove the efficacy of the proposed work are also highlighted in this section. Finally, Section 5 concludes the paper by stating the findings and future directions of the proposed work.

2. Related work

The recent studies related to the proposed work are organized into two subsections. The first subsection presents a survey of the feature selection (FS) process and its different types for the automatic text classification process. The subsecond section describes various studies related to depression detection using social media.

2.1. Feature selection. Dimensionality reduction plays a significant role in the automatic text classification process. Two processes of dimensionality reduction are feature extraction and feature selection. Feature extraction is the process of creating a brand new set of features from the original feature set which captures the required information. Feature selection is the process of selecting an optimal feature subset from the original feature space based on some criteria such as accuracy, error rate, etc. Deng *et al*. (2019) provided a survey of various feature selection methods. The main difference is that feature extraction generates a new feature set and feature selection retains the original feature subset.

Selecting an optimal feature subset from the original set of features is characterized as an NP-hard problem (Kowal *et al*., 2018). The key objective of the feature selection process is to reduce the dimension of the original corpus with a minimum number of features and maximum accuracy. Hence, the feature selection problem is considered as a multi-objective problem. Feature selection is a preeminent process in machine learning applications (Rajalakshmi and Aravindan, 2018) as it serves as an elementary procedure to choose the variables or attributes to direct the model learning process in a more effective manner. The machine learning model proposed by Durgalakshmi and Vijayakumar (2020) used a feature selection technique to identify breast cancer in the WDBC (Wisconsin Diagnostic Breast Cancer) database. By the some token, a similarity-based feature selection model is proposed by Zhu *et al*. (2019) for an unsupervised machine learning process.

2.1.1. Types of feature selection methods. Feature selection algorithms are grouped into three different categories: filter methods, wrapper methods, and hybrid methods.

Filter methods. Filter based FS methods concentrate on the properties of data, rather than the learning algorithms. Filter methods are less prone to over-fitting. They calculate the score for each feature based on certain formulas. Then the features are sorted based on the calculated scores and select the top features for classification. There are many filter-based FS approaches proposed in the literature. Some of the existing feature selection methods are document frequency (DF) (Li *et al*., 2015), balanced accuracy (ACC2) (Asim *et al*., 2017), information gain (IG) (Gao *et al*., 2014), the Gini index (GINI) (Sanasam *et al*., 2010) and the normalized difference measure (NDM) (Rehman *et al*., 2017). One of the recently proposed filter-based methods is the extensive feature selector (EFS). This method combines both class-level and corpus-level probabilities to select more informative and distinct features when compared to other filter based methods.

Wrapper methods. Wrapper based feature selection methods evaluate the candidate feature subsets using a machine learning algorithm and select an optimal set of features. The recent interest in wrapper methods is related to metaheuristic algorithms (Xue *et al*., 2016; Prachi *et al*., 2021) due to their simplicity and great performance in feature selection. Numerous meta-heuristic algorithms have been proposed so far and some of the recent findings are the grey wolf optimization algorithm (GWO) (Emary *et al*., 2016b), the ant lion optimization algorithm (ALO) (Emary *et al*., 2016a), the Jaya optimization algorithm (JO) (Rao, 2016), the whale optimization algorithm (WOA) (Hussien *et al*., 2020), and the gaining-sharing knowledge-based optimization algorithm (GSK) (Mohamed *et al*., 2020). Meta-heuristic algorithms have a wider scope of applications in various domains. For instance, a tabu search (TS) and a genetic algorithm (GA) are applied to solve unrelated machine scheduling problems (Wang *et al*., 2020) with the criteria of late work jobs. Wrapper methods have high computational complexity when compared with filter methods.

Hybrid methods. Hybrid methods are combination of both filter and wrapper methods. They highlight the efficiency of the filter methods and the accuracy of the wrapper methods. There are two types of hybrid algorithms: a two-stage algorithm and an embedded algorithm. In the two-stage algorithm (Peng *et al*., 2005; Uysal, 2018), the filter and wrapper methods are treated as two different steps. In the embedded model, either the wrapper method is embedded in the filter method (Unler *et al*., 2011) or the filter method is embedded in the wrapper method. Moradi and Gholampour (2016) embedded a local search strategy into particle swarm optimization (PSO) for feature selection in the machine learning model. The NDM-BJO hybrid feature selection method is proposed by Thirumoorthy and Muneeswaran (2020) for text classification. A hybrid feature selection algorithm for ensemble classification is proposed by Moorthy and Gandhi (2019). Hybridization can be applied to any real world optimization problems. Currently, hybridization with evolutionary search strategies is one of the subject of interests in the text feature selection research area. Adhering to that fact, the proposed work integrates EFS and pBGSK techniques to select relevant and optimal numbers of features for classifying short text data.

2.2. Depression detection through social media. The information shared by depressed individuals on social media or the Internet paves the way to develop intelligent systems to deal with depression prediction problems. Many researches have developed such intelligent systems to detect depression through social media. A study of anxiety disorder using Reddit posts is presented by Shen and Rudzicz (2017). This approach distinguishes anxiety related posts from typical posts using features extracted based on different feature generation models like the vector space embeddings, latent Dirichlet allocation (LDA) topic modelling, linguistic inquiry and word count (LIWC) features and an N-gram language model. The accuracy of feature vectors obtained from Reddit texts is higher when compared with Twitter texts. This result is likely due to the fact that the lengths of Reddit texts are larger than Twitter texts. Therefore, this model works well for larger texts rather than shorter text sequences.

Husseini Orabi *et al*. (2018) presented a neural network architecture to identify depressed users from their social media posts. To learn a better feature representation, the features are extracted using optimized word embedding vectors. Then, the models are trained based on four neural network models. Traditional machine learning models (Islam *et al*., 2018) are also used to perform depression analysis based on four different types of features extracted from Facebook data. The early depression detection (EDD) problem is addressed by Trotzek *et al*. (2018). This work developed a convolutional neural network model based on different word embeddings using social media messages. The authors also proposed a modified early risk detection error (ERDE) metric to evaluate the model.

A general supervised learning framework to handle early risk detection (ERD) problems on social media is developed by Burdisso *et al*. (2019). In this work, a text classification model is presented based on three aspects called SS3 (Sequential S3). The relationship between a user's linguistic metadata and depression is investigated by Tadesse *et al*. (2019). This study demonstrated the effectiveness of combined features in classification of depression related posts from the Reddit forum. The research work presented by Thorstad and Wolff (2019) examined whether the future occurrence of several types of a mental illness can be predicted using people's everyday language. A socially mediated patient portal (SMPP) application was developed by Hussain *et al*. (2019) to detect the features to characterize depressed and nondepressed Facebook users. To detect depression among college students, a deep integrated support vector algorithm is designed by Ding *et al*. (2020). The algorithm uses Sina Weibo (a Chinese microblogging website) data and the text features are extracted using the deep neural networks. Using Twitter data, generalized text based depression detection adapting various supervised machine classifiers was investigated by Chiong *et al*. (2021). It is to be noted that combining social media data and ML models gives more effective results in identifying depression markers.

The surveys conducted by William and Suhartono (2021) as well as Babu and Kanaga (2022) show that the depression detection based on artificial intelligence (AI) techniques utilizing data available in social media is one of the promising research fields. In this context, the proposed work introduces a novel hybrid algorithm with an evolutionary approach to identify depression

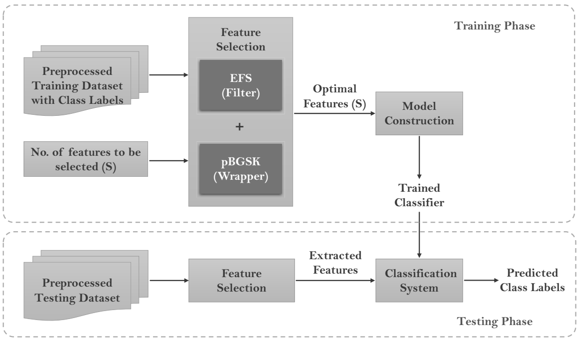


Fig. 1. Overall architecture of the proposed hybrid feature selection approach.

related content using text sequences. The feature selection process embeds the filter algorithm into the wrapper algorithm to select the depression related features with superior classification performance in terms of both accuracy and F_β scores. The EFS is used to filter and rank the features and the pBGSK is used as the wrapper algorithm to select an optimal subset of features.

3. Proposed model

In the proposed work, a metaheuristic approach is used to select an optimal set of features for a depression classification system using text sequences. The overall architecture of the proposed feature selection approach (EFS-pBGSK) is shown in Fig. 1. This section discusses the different modules of the proposed model along with some preliminaries of the text feature selection problem, the EFS method and the pBGSK algorithm.

3.1. Text feature selection. A text corpus comprises text documents/sequences, and the text classification problem is to categorize the document into an appropriate category by assigning the relevant class label(s). Feature selection is an inevitable task in the text classification process. It is the process of selecting an optimal feature subset from the original feature space of the training set. Then only this subset is used as features for classification. Feature selection reduces the dimension of the feature space by eliminating the redundant, irrelevant and noise features and selecting the optimal features, thereby improving the accuracy and efficiency of the classifier.

3.1.1. Document term matrix. The proposed work applies the document term matrix (DTM) for representing the text corpus. The DTM is a conversion technique used to transform the text data into mathematical matrices, where the rows denote the instances in the corpus, the columns denote the word/term features of the corpus and the cell values represent the term frequencies.

3.2. Extensive feature selector. To obtain a more distinct and explicit set of features, the EFS method (Parlak and Uysal, 2021) is exerted in the proposed system. It calculates the significance of features based on both class-level and corpus-level probabilities. The mathematical formulation of the EFS method is given as

$$\mathrm{EFS}(t) = \sum_{k=1}^{l} \mathrm{EFS}_{\text{class-level}} \cdot \mathrm{EFS}_{\text{corpus-level}}, \tag{1}$$

$$\mathrm{EFS}_{\text{class-level}} = \frac{P(t|C_k)}{P(\bar{t}|C_k) + P(t|\bar{C}_k) + 1}, \tag{2}$$

$$\mathrm{EFS}_{\text{corpus-level}} = \frac{P(C_k|t)}{P(\bar{C}_k|t) + P(C_k|\bar{t}) + 1}. \tag{3}$$

Equation (2) is applied to compute the class level(C_k) score of each feature/term depending on t based on the conditional probability $P(t|C_k)$. The corpus/dataset level score of each feature is assessed using Eqn. (3) with the conditional probability $P(C_k|t)$. Finally, the class-level and dataset-level scores are multiplied and summed up to obtain the final score of each feature (EFS(t)) as given in Eqn. (1).

The variable C_k in the aforementioned equations represents the class label and $k = 1, 2, \dots, l$ (l is the number of class labels in the dataset). The EFS score value ranges from 0.0 to 1.0. The maximum score value 1.0 indicates that the term t exists only in one class across all the documents and reveals the uniqueness of the feature. The notation used in the above equations is given in Table 1.

3.3. Gaining-sharing knowledge-based algorithm. The gaining-sharing knowledge-based (GSK) optimization algorithm (Mohamed *et al.*, 2020) is a metaheuristic approach inspired by the human behaviour. The algorithm mimics the human nature of gaining and sharing knowledge during their entire lifespan. The concept of the GSK algorithm depends on two phases:

(i) junior gaining and sharing knowledge phase (beginners–intermediate or early middle phase),

(ii) senior gaining and sharing knowledge phase (intermediate–experts or middle later phase).

In the junior phase, the beginner gains and shares knowledge with only known family members, neighbors or relatives. On the contrary, in the senior phase, the individual can able to identify good and bad contacts

Table 1. Notation used in the EFS-pBGSK method.

Notation	Value	Description
a	count(t, C_k)	Number of documents containing term t in class C_k
b	count$(t, \bar{C}_k)$	Number of documents containing term t in other classes $\bar{C}_k$
c	count$(\bar{t}, C_k)$	Number of documents not containing term t in class C_k
d	count$(\bar{t}, \bar{C}_k)$	Number of documents not containing term t in other classes $\bar{C}_k$
$P(t\|C_k)$	$\frac{a}{a+c}$	Probability of term t when class C_k is present
$P(\bar{t}\|C_k)$	$\frac{c}{a+c}$	Probability that term t is absent when class C_k is present
$P(t\|\bar{C}_k)$	$\frac{b}{b+d}$	Probability of term t when other classes($\bar{C}_k$) are present
$P(C_k\|t)$	$\frac{a}{a+b}$	Probability of class C_k when term t is present
$P(\bar{C}_k\|t)$	$\frac{b}{a+b}$	Probability that class C_k is absent when term t is present
$P(C_k\|\bar{t})$	$\frac{c}{c+d}$	Probability of class C_k when term t is absent

and gains and shares knowledge from a large network of suitable friends, colleagues or experts. The search process of the GSK algorithm retains the equilibrium between the local exploitation and the global exploration tendency with the help of the distinct junior and senior phases. This is the key advantage of adapting the GSK algorithm over recently proposed meta-heuristic algorithms. As both junior and senior phases update the knowledge vector independently, they can be executed simultaneously using different cores. Accordingly, the GSK algorithm supports parallel programming which is not feasible in other recent algorithms including red fox optimization (Połap and Woźniak, 2021) and black widow optimization (Hayyolalam and Kazem, 2020) algorithms.

3.4. Binary GSK (BGSK) algorithm. The binary GSK metaheuristic approach (Agrawal *et al.*, 2021) is introduced to deal with problems in binary intervals and is based on the conventional GSK algorithm with knowledge factor $k_f = 1$. This subsection presents the binary initialization, the dimension of the two phases and the working methodology of both the phases (junior and senior gaining-sharing knowledge phases) in the binary space.

3.4.1. Solution encoding. Consider T as a text corpus with d number of text documents, l the number of class labels and w the number of features (words or terms). Let W be the set of real numbers that represents all the w features. The feature selection mechanism has to select an optimal set of features from the original feature set W by optimizing the objective function $f(X)$. The binary encoding strategy is applied to encode the solution(X) as follows:

$$X = \{(x_{i1}, x_{i2}, \dots, x_{iw}) : x_{ij} \in \{0,1\};\ i = 1, 2, \dots, d\}.$$

In the solution vector, the value of $x_{ij} = 1$ indicates that the j-th feature is selected and $x_{ij} = 0$ indicates that the j-th feature is not selected.

3.4.2. Candidate initialization. In the proposed feature selection algorithm, NP represents the total number of individuals/persons. Every individual in the population is denoted by x_t, where $t = \{1, 2, \dots, \text{NP}\}$. The knowledge gained by an individual is represented as $x_{tk} = (x_{t1}, x_{t2}, \dots, x_{td})$, where d is the dimension of the feature space. The corresponding objective function values are denoted as f_t. The individual vector is a binary vector, i.e., the values are either 0 or 1. The value 1 indicates that the feature at the corresponding index is selected and 0 indicates that the feature at the corresponding index is not selected. Here x_{tj} represents the value of feature j in the individual vector j at the t-th iteration. The initial candidate solutions(x^0_{tj}) are made binary using (Agrawal *et al.*, 2021)

$$x^0_{tj} = \text{round}(\text{rand}(0, 1)), \tag{4}$$

where the round operator approximates the continuous random value by the nearest binary value (0 or 1). The dimensions of the junior and senior phases are evaluated using the formula given by Agrawal *et al.* (2021).

3.4.3. Junior gaining and sharing knowledge phase. In this phase, the individuals are arranged in the order of importance based on the values of the objective function. Then, the nearest best (x_{t-1}) and worst (x_{t+1}) individuals are chosen to gain knowledge. Then a randomly selected individual (x_R) is used to share the knowledge. The new solutions (x^{new}_{tk}) are updated based on the following two cases:

Case 1. $f(x_R) < f(x_t)$

$$x^{\text{new}}_{tk} = \begin{cases} x_R & \text{if } x_{t-1} = x_{t+1}, \\ x_{t-1} & \text{if } x_{t-1} \neq x_{t+1}. \end{cases} \tag{5}$$

Case 2. $f(x_R) \geq f(x_t)$

$$x_{tk}^{\text{new}} = \begin{cases} x_{t-1} & \text{if } x_{t-1} \neq x_{t+1} = x_R, \\ x_t & \text{otherwise.} \end{cases} \tag{6}$$

3.4.4. Senior gaining and sharing knowledge phase. In this phase, the individuals are arranged in the order of importance based on the values of the objective function. Then, the individuals are grouped into best (x_{pbest}), middle (x_{middle}) and worst (x_{pworst}) categories. The individual gains knowledge from two randomly chosen individuals from the best and worst categories. The middle individual is used to share the knowledge. The new solutions (x_{tk}^{new}) are updated based on two cases:

Case 1. $f(x_{\text{middle}}) < f(x_t)$

$$x_{tk}^{\text{new}} = \begin{cases} x_{\text{middle}} & \text{if } x_{\text{pbest}} = x_{\text{pworst}}, \\ x_{\text{pbest}} & \text{if } x_{\text{pbest}} \neq x_{\text{pworst}}. \end{cases} \tag{7}$$

Case 2. $f(x_{\text{middle}}) \geq f(x_t)$

$$x_{tk}^{\text{new}} = \begin{cases} x_{\text{pbest}} & \text{if } x_{\text{pbest}} \neq x_{\text{pworst}} = x_{\text{middle}}, \\ x_t & \text{otherwise.} \end{cases} \tag{8}$$

3.5. EFS-pBGSK based feature selection algorithm. The pBGSK optimization algorithm (Agrawal *et al.*, 2021) is a unique variant of the BGSK algorithm associated with population reduction strategy. This approach is used to gradually reduce the population size to improve the algorithm performance. The pBGSK algorithm selects a varying number of feature subsets for every iteration. To amend the feature subset with the required/fixed number of features, the extensive feature selector (EFS) measure discussed in Section 3.2 is employed. Using this approach, when the feature vector goes beyond the search space, the remaining features with top ranked EFS scores will be selected. When the feature vector suffers with the least number of features, the excluded features with high EFS scores can be included.

The pseudocode for the text feature selection for depression classification using the proposed EFS-pBGSK approach is given in Algorithm 1. The algorithm first generates the initial candidate population (x_t) using Eqn. (4) as mentioned in Line 1. The next step is to ensure that the number of features selected from the initial population matches the required number of features (ns) as shown in Lines 2–5. The function `SelectedFeatures()` in Line 3 selected the features whose binary value is 1 and the function `count()` counts the number of features selected. If the count does not match the value ns, then the feature set will be updated based on the EFS score of the features.

Algorithm 1. Optimal feature selection algorithm using EFS-pBGSK.

Require: D_{train}: DTM of preprocessed training dataset,
D_{test}: DTM of preprocessed testing dataset,
$\text{NP}_{\min}$, $\text{NP}_{\max}$: minimum and maximum population counts,
$G_{\max}$: maximum number of iterations,
w: dimension of feature space in dataset,
ns: number of features to be selected

1: Generate the initial population of individuals x_t
2: **for** $t = 1 : \text{NP}$ **do**
3:   **if** count(SelectedFeatures(x_t)) $\neq ns$ **then**
4:     Mutate the knowledge vector x_t using EFS scores of features {cf. Eqn. (1)}
5:   **end if**
6:   $x_t^{\text{fitness}} = \text{calculateFitness}(x_t, D_{\text{train}}, D_{\text{test}})$
7: **end for**
8: $x_{\text{best}} = \text{findBestIndividual}()$ {Select a locally best individual based on fitness value}
9: $k_r = \text{rand}$; $G = 1$; $\text{NFE} = G \times \text{NP}_{\max}$
{Finding the optimal individual}
10: **while** NFE < MaxNFE **do**
11:   $w_{\text{junior}} = w \times \left(1 - \frac{\text{NFE}}{\text{MaxNFE}}\right)^{k_r}$
12:   $w_{\text{senior}} = w - w_{\text{junior}}$
  {Generating new knowledge vector for individuals}
13:   **for** $t = 1 : NP$ **do**
14:     **for** $y = 1 : w_{\text{junior}}$ **do**
15:       Apply binary junior GSK phase (cf. Section 3.4.3)
16:     **end for**
17:     **for** $y = w_{\text{junior}} + 1 : w_{\text{senior}}$ **do**
18:       Apply binary senior GSK phase (cf. Section 3.4.4)
19:     **end for**
20:     $x_t^{\text{fitness}} = \text{calculateFitness}(x_t, D_{\text{train}}, D_{\text{test}})$
21:   **end for**
  {Updating population count and NFE for next iteration}
22:   $G{+}{+}$
23:   $\text{NP}_G = (\text{NP}_{\min} - \text{NP}_{\max}) \times \left(\dfrac{\text{NFE}}{\text{maxNFE}}\right) + \text{NP}_{\max}$
24:   $\text{NFE} = G \times \text{round}(NP_G)$
25:   **if** count(SelectedFeatures(x_t)) $\neq ns$ **then**
26:     Mutate knowledge vector x_t using EFS scores of features, {cf. Eqn. (1)}
27:   **end if**
28:   $x_{\text{best}} = \text{findBestIndividual}()$
29: **end while**
30: OptFeatures = selectFeatures(x_{best}) {Extract an optimal set of features from the global best individual}
31: **return** OptFeatures

Then, the fitness score of the individual knowledge vector is calculated using Eqn. (9) as given in Line 6. The individual with a best fitness value(less error rate) is selected in Line 8. In Line 9, the parameters required for the optimization process are initialized. The knowledge rate (k_r) is initialized with a random number (rand), the iteration number (G) is set as 1 and the number of function evaluations (NFE) is calculated as $G \times \mathrm{NP}_{\max}$. The iteration process of EFS-pBGSK optimization is illustrated in Lines 10–29. Lines 11 and 12 depict the dimension computation for junior (w_{junior}) and senior(w_{senior}) phases, respectively. Lines 13–21 describe the process of new knowledge vector generation for every individual in the population. Lines 22–24 update the parameters for the next iteration. Lines 25–27 update the selected feature set and the local best individual is found in Line 28. After the whole iteration process, the optimal feature set is extracted from the global best individual as shown in Line 30.

3.5.1. Fitness function. The primary objectives of selecting an optimal feature subset are to minimize the feature dimension and to maximize the classification accuracy simultaneously. Hence, this multi-objective problem is expressed via a single objective function (Z) (Agrawal *et al.*, 2021)

$$\min Z(F) = \gamma_1 \cdot E_{\text{rate}} + \gamma_2 \cdot \frac{\text{number of selected features}}{\text{total number of features}}, \tag{9}$$

where F indicates the fitness function, E_{rate} denotes the classification error rate, γ_1 and γ_2 are two fitness parameters that are symmetric with respect to the subset length and can be computed as $\gamma_1 \in [0, 1]$ and $\gamma_2 = 1 - \gamma_1$.

3.5.2. Termination criterion. The optimization approach follows the iterative procedure which needs a stopping criterion to end the process. The termination point depends on various factors including the maximum number of generations, the convergence rate and so on. The termination criterion of the proposed algorithm is based on the maximum number of function evaluations (MaxNFE). It is obtained by multiplying the maximum number of populations by the maximum number of generations (MaxNFE = $\mathrm{NP}_{\max} \times G_{\max}$).

3.5.3. Parameter settings. The hyperparameter values of the pBGSK algorithm are fixed based on the trial-and-error method to achieve a better text depression classification performance. Table 2 provides the values initialized for the parameters used in the EFS-pBGSK algorithm for text feature selection.

4. Experimental settings

4.1. Datasets. In order to test the performance of the proposed feature selection approach, two depression detection datasets are used: *Sentiment 140* and *Suicide and Depression Detection*. These datasets are publicly available in the Kaggle repository. The datasets mainly contribute to the depression detection problem, since the posts are extracted using hashtags related to the depression and suicide. Therefore, the posts reveal the markers of depression and suicide which is very helpful in training the proposed model. *Sentiment 140* (Senti140) comprises 1.6 million tweets extracted using Twitter API. The tweets are labelled as positive or negative sentiments. *Suicide and Depression Detection* (SDD) comprises 2.3 million Reddit posts collected from “depression” and “SuicideWatch” subreddits. The training-testing split up of the dataset is set as 75%–25%, respectively. From *Sentiment 140*, 1.2 million tweets are used for training the model and 0.4 million tweets are used for model testing. Similarly, from the SDD dataset, 172,500 Reddit posts are used for training and 57,500 Reddit posts are used for testing.

4.2. Preprocessing. Text preprocessing is an integral process in text classification as it enhances the information quality and performance of the model. The preprocessing of the social media posts comprises lowercasing, removal of URL, digits, stop words and punctuation, tokenization and stemming. Further, the dataset is pruned to remove the rare features and more frequent features which lead to overfitting and reduce the model accuracy. In this work, the dataset is pruned to eliminate the features that are present in more than 0.75% (more frequent) of the dataset or present in less than 0.25% (very rare) of the dataset.

4.3. Classifiers. Two widely used classification algorithms are employed to verify the performance of the classification based on the proposed methodology: the support vector machines (SVMs) and the naive Bayes classifier (NB). The SVM (Suthaharan, 2016; Derek and David, 2020) is a well-known classification algorithm used in machine learning. The objective of the SVM is to find the best decision boundary, also called the hyperplane, which can separate the n-dimensional feature space into categories. The SVM selects many points/vectors in finding the hyperplane and all these points are called support vectors. The SVMs can be divided into two types based on the kernel functions: a linear SVM and a nonlinear SVM. In our experiments, the linear SVM function from `scikit-learn` 0.24.2 library in Python is used with its default parameter settings.

The naive Bayes classifier (Chen *et al.*, 2009) is a simple and effective classifier which can be applied for

Table 2. Parameter initialization.

Parameter	Description	Value
$NP_{\max}$	maximum population count	50
$NP_{\min}$	minimum population count	5
$G_{\max}$	maximum number of iterations	15
MaxNFE	maximum number of function evaluations	$\mathrm{NP}_{\max} \times G_{\max}$
γ_1	parameter in fitness function	0.99
γ_2	parameter in fitness function	0.01

high dimensional data to build fast prediction models. The working concept of the NB is based on the Bayes Theorem and the assumption of conditional independence. The primary idea is to use the joint probabilities of terms and classes to determine the class of a given document. Given the document $d = (t_1, t_2, \dots, t_n)$ and the set of class labels C_k, the class label can be predicted as follows:

$$\text{label}(d) = \max_{i=1,2,\dots,k} P(C_i) \prod_{j=1}^{n} P(t_j|C_i). \tag{10}$$

4.4. Performance metrics. The experimental performance was evaluated with the help of four different metrics predominant for evaluating classification models: accuracy, F1 score, MAE and log-loss. Accuracy is the ratio of correctly classified documents,

$$\begin{aligned}&\text{accuracy}\\&\quad= \frac{\text{number of correctly classified documents}}{\text{total number of documents}}.\end{aligned} \tag{11}$$

The F-score is the weighted average of precision and recall, giving equal importance to both precision and recall. The F_β score is an F-score generalization with β as the configuration parameter which is a positive real number. The case of $\beta = 1$ is same as that of the F-score, $\beta < 1$ assigns more weighting to precision than recall and $\beta > 1$ assigns more weighting to recall than precision. For the assessment, the beta values are set as $\beta = 0.5, 1, 2$ to test all the three cases. The score can be computed using the formula

$$F_\beta = (1 + \beta^2) \times \frac{\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}. \tag{12}$$

The MAE evaluates the error rate of a model by means of averaging the absolute error difference between the predicted label and actual label of all the instances.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{m} |p_i - a_i|, \tag{13}$$

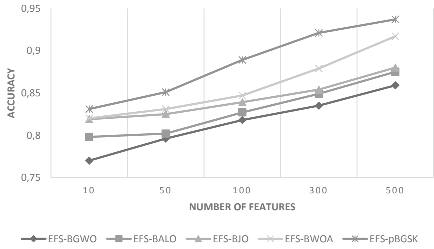
where m, p_i and a_i represent the number of instances (posts), the predicted label and the actual label, respectively.

The log-loss evaluates the model based on the maximum log likelihood probability estimation. It implies the proximity between the predicted probability and the actual ground truth value. The lower the log-loss value, the closer the predicted probability is to the actual value. We have

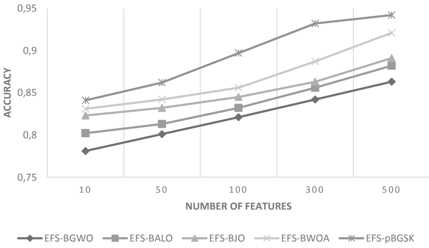
$$\begin{aligned}&\text{log-loss}\\&\quad= -\frac{1}{N} \sum_{i=1}^{m} a_i \log(p_i) + (1 - a_i) \log(1 - p_i).\end{aligned} \tag{14}$$

4.5. Results and a discussion. The main objective of the proposed FS approach is to extract depression related features from the text sequences. This section presents the various performance comparisons of the proposed embedded feature selection approach (EFS-pBGSK) with other existing algorithms to prove the efficacy of the proposed work. The performance of the proposed hybrid algorithm is compared with other evolutionary algorithms such as Binary GWO, Binary ALO, Binary JO and Binary WOA combined with the EFS approach. The experiments are carried out in a computer equipped with AMD Ryzen 5 3500U with Radeon Vega Mobile Gfx 2.10 GHz and 8.00 GB RAM. Python 3.7.6 is used for the implementation.

Various comparison graphs are plotted to justify the performance of the proposed EFS-pBGSK approach on previously discussed datasets. Figures 2 and 4 show the accuracy graphs plotted to compare the performance of the proposed EFS-pBGSK method with other feature selection methods using NB and SVM classifiers on the two datasets, respectively. In all the accuracy graphs, the horizontal axis indicates the increasing number of features and the vertical axis indicates the performance of the classifier in terms of the accuracy. The bar graphs shown in Figs. 3 and 5 denote the MAE and log-loss comparison between the proposed hybrid model and other hybrid models. Comparisons $F_{0.5}$, F_1 and F_2 scores of different FS methods are included in Tables 3 and 4. In all the tables, the values $10, 50, 100, 300$ and 500 represent the selected numbers of features. Figure 6 illustrates the consistency of the proposed work in terms of accuracy using box plots. The comparison shows that the proposed embedded feature selection approach outperforms the other hybrid algorithms.

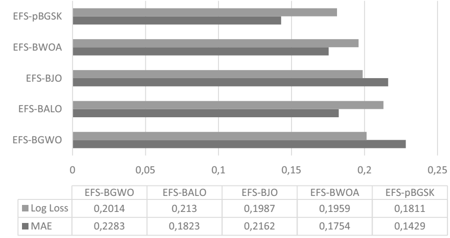


(a) NB classifier

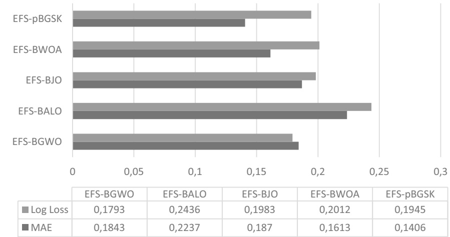


(b) SVM classifier

Fig. 2. Accuracy comparison for the Sentiment 140 dataset.



(a) NB classifier



(b) SVM classifier

Fig. 3. MAE and log-loss comparison for the Sentiment 140 dataset.

4.5.1. Performance comparison on the Sentiment 140 dataset. The tweets in the Sentiment 140 dataset are classified into two classes: depressive posts (labelled as 0 or negative) and non-depressive posts (labelled as 1 or positive). The classification accuracy on Sentiment 140 using NB and SVM classifiers based on various embedded FS approaches depicted in Fig. 2. Using the NB classifier, the proposed methodology attains the classification accuracies of 0.831, 0.851, 0.889, 0.921 and 0.937 while selecting feature subsets of sizes 10, 50, 100, 300 and 500, respectively. Similarly, using the SVM classifier, the proposed methodology achieves the classification accuracies of 0.841, 0.862, 0.897, 0.932 and 0.942 while selecting feature subsets of sizes 10, 50, 100, 300 and 500, respectively. From the obtained accuracy values, it can be seen that the SVM classifier works quite better than the NB classifier for the Sentiment 140 dataset. The accuracy curve trends of both the classifiers in Fig. 2 exhibit the superior performance of the proposed approach when compared with the other blended approaches. From Fig. 3, it is obvious that the error rate of the proposed model is comparatively low, and the depression prediction probability of the model is very close to the actual label.

Table 3 illustrates the comparison of F_β score values of the present method for $\beta = 0.5, 1, 2$. The $F_{0.5}$ score values show that the proposed algorithm assigns more relevant class labels to the data instances and yields more quality with maximum scores of 0.915 and 0.921 using the NB and SVM classifiers, respectively. The F_2 scores values indicate that the proposed algorithm classifies most of the data correctly with maximum scores of 0.928 and 0.931 using the NB and SVM classifiers, respectively. The scores indicate that the proposed algorithm achieves higher recall than precision. Considering both precision and recall equally, the proposed work attains maximum F_1 score values of 0.921 and 0.929 using the NB and SVM classifiers, respectively. From the overall analysis on the Sentiment 140 dataset, it is evident that the performance of the proposed feature selection model is remarkable when compared with the other optimization algorithms embedded with the EFS approach.

4.5.2. Performance comparison on the SDD dataset. In the SDD dataset, the Reddit posts are labelled with one of the two unique categories: suicide and non-suicide. Figure 4 presents the graph of classification accuracy obtained for the SDD Reddit post classification using various hybrid approaches based on the NB and SVM classifiers. For varying the size of the feature subset as 10, 50, 100, 300 and 500, the NB classifier gains the accuracy of 0.845, 0.883, 0.929, 0.937 and 0.959, respectively. Similarly, the SVM classifier produces accuracy scores of 0.854, 0.897, 0.935, 0.941, and 0.962 for the different feature subset sizes. From the observations, it is quite evident that the SVM classifier works well for both the depression detection datasets when compared with the NB classifier. The trending accuracy curves of both the classifiers in Fig. 4 makes the present methodology acquire greater performance in comparison with the other combined state-of-art optimization algorithms. The error rate observed from the graphs in Fig. 5 reveals that the prediction possibility of the proposed model is far better when compared with other models under consideration.

The comparison of $F_\beta(\beta = 0.5, 1, 2)$ scores of

Table 3. F_β (β = 0.5, 1 and 2) score comparison for depression detection on the Sentiment140 dataset using an NB classifier (a) and an SVM classifier (b).

F_β	Classifier	FS Methods	10	50	100	300	500
$F_{0.5}$	NB	EFS-BGWO	0.749	0.765	0.783	0.814	0.823
		EFS-BALO	0.752	0.778	0.794	0.814	0.838
		EFS-BJO	0.771	0.794	0.816	0.836	0.859
		EFS-BWOA	0.789	0.802	0.825	0.847	0.871
		EFS-pBGSK	0.801	0.825	0.863	0.897	0.915
	SVM	EFS-BGWO	0.754	0.773	0.798	0.823	0.841
		EFS-BALO	0.768	0.781	0.802	0.826	0.849
		EFS-BJO	0.782	0.804	0.829	0.847	0.867
		EFS-BWOA	0.811	0.819	0.829	0.856	0.889
		EFS-pBGSK	0.817	0.831	0.869	0.902	0.921
F_1	NB	EFS-BGWO	0.751	0.774	0.792	0.821	0.831
		EFS-BALO	0.761	0.784	0.796	0.816	0.849
		EFS-BJO	0.782	0.801	0.823	0.840	0.866
		EFS-BWOA	0.791	0.815	0.829	0.858	0.894
		EFS-pBGSK	0.816	0.829	0.871	0.904	0.921
	SVM	EFS-BGWO	0.762	0.781	0.798	0.827	0.850
		EFS-BALO	0.771	0.793	0.813	0.834	0.856
		EFS-BJO	0.789	0.812	0.831	0.856	0.871
		EFS-BWOA	0.819	0.826	0.837	0.862	0.892
		EFS-pBGSK	0.836	0.859	0.873	0.918	0.929
F_2	NB	EFS-BGWO	0.759	0.781	0.798	0.827	0.845
		EFS-BALO	0.768	0.790	0.801	0.827	0.852
		EFS-BJO	0.789	0.811	0.828	0.848	0.872
		EFS-BWOA	0.797	0.821	0.839	0.862	0.901
		EFS-pBGSK	0.822	0.847	0.878	0.912	0.928
	SVM	EFS-BGWO	0.768	0.789	0.812	0.826	0.857
		EFS-BALO	0.778	0.797	0.819	0.840	0.861
		EFS-BJO	0.791	0.818	0.837	0.862	0.889
		EFS-BWOA	0.823	0.834	0.842	0.869	0.912
		EFS-pBGSK	0.831	0.845	0.882	0.925	0.931

different embedded methods with the currently presented work on the SDD dataset is displayed in Table 4. The proposed feature selection model obtains a maximum $F_{0.5}$ score of 0.935 and 0.949 using NB and SVM classifiers, respectively, which conveys that the data instances are tagged with more relevant class labels featuring better quality. The maximum F_2 scores 0.951 and 0.957 gained respectively using NB and SVM classifiers reveal that most posts are classified correctly employing the proposed method. When comparing the $F_{0.5}$ and F_2 scores, it can be found that the model has higher recall than precision. On giving equal importance to both precision and recall, the maximum F_1 scores of 0.942 and 0.952 can be realized using NB and SVM, respectively. The resulting values signify that the proposed algorithm realizes better classification quality in terms of both precision and recall. Additionally, the proposed multi-objective EFS-pBGSK algorithm yields impressive outcomes compared with the other multi-objective algorithms.

The main motive behind adopting the optimization algorithm in the proposed feature selection approach is to select the minimal number of features required for classification. Without optimization, we can achieve the reasonable performance with the high number of features which increases the time and space consumption. The novel pBGSK optimization algorithm drastically reduces the percentage of features needed for the classification and simultaneously enhances the performance of the classifier. Since the SVM classifier works better for both the datasets, it is used for the final optimization evaluation in Table 5. This table justifies the importance and the performance of the proposed hybrid feature selection model for depression detection using the SVM classifier. The values in boldface indicate the overall performance improvement achieved by the proposed embedded feature selection model.

Table 4. F_β (β = 0.5, 1 and 2) score comparison for depression detection on the SDD dataset using an NB classifier (a) and an SVM classifier (b).

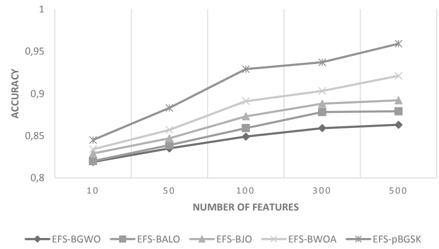
F_β	Classifier	FS methods	10	50	100	300	500
$F_{0.5}$	NB	EFS-BGWO	0.765	0.782	0.794	0.813	0.842
		EFS-BALO	0.749	0.787	0.802	0.818	0.835
		EFS-BJO	0.769	0.798	0.826	0.847	0.860
		EFS-BWOA	0.778	0.813	0.836	0.852	0.894
		EFS-pBGSK	0.817	0.851	0.879	0.904	0.935
	SVM	EFS-BGWO	0.772	0.791	0.807	0.824	0.853
		EFS-BALO	0.754	0.790	0.814	0.826	0.849
		EFS-BJO	0.781	0.805	0.832	0.855	0.873
		EFS-BWOA	0.774	0.823	0.839	0.861	0.908
		EFS-pBGSK	0.824	0.865	0.905	0.924	0.949
F_1	NB	EFS-BGWO	0.769	0.785	0.792	0.817	0.846
		EFS-BALO	0.751	0.792	0.809	0.822	0.839
		EFS-BJO	0.773	0.804	0.837	0.851	0.862
		EFS-BWOA	0.781	0.817	0.839	0.857	0.903
		EFS-pBGSK	0.820	0.867	0.887	0.915	0.942
	SVM	EFS-BGWO	0.781	0.796	0.811	0.834	0.853
		EFS-BALO	0.758	0.796	0.819	0.831	0.851
		EFS-BJO	0.787	0.811	0.839	0.862	0.884
		EFS-BWOA	0.782	0.833	0.845	0.869	0.915
		EFS-pBGSK	0.835	0.871	0.917	0.938	0.952
F_2	NB	EFS-BGWO	0.772	0.791	0.797	0.824	0.846
		EFS-BALO	0.757	0.795	0.812	0.826	0.843
		EFS-BJO	0.771	0.805	0.839	0.859	0.867
		EFS-BWOA	0.785	0.819	0.846	0.861	0.909
		EFS-pBGSK	0.837	0.872	0.895	0.923	0.951
	SVM	EFS-BGWO	0.785	0.802	0.819	0.838	0.862
		EFS-BALO	0.763	0.798	0.905	0.839	0.857
		EFS-BJO	0.792	0.817	0.844	0.869	0.895
		EFS-BWOA	0.786	0.839	0.847	0.873	0.917
		EFS-pBGSK	0.839	0.877	0.921	0.945	0.957

4.5.3. Consistency and convergence comparison. Figure 6 presents the box plots of the classification accuracy obtained from 20 runs of the proposed feature selection algorithm and other algorithms for the two datasets. The box plot of the proposed algorithm shows that there is less dispersion in the accuracy, which proves that the consistent performance is gained from 20 runs. The proposed algorithm could achieve this consistent performance with a lower number of features using a simple and feasible feature selection model without a highly complex and sophisticated model. The convergence curves of different algorithms for the Sentiment 140 and SDD datasets are illustrated in Fig. 7. Irrespective of the size of the feature subset, the proposed algorithm can classify the depression related texts with a minimal set of features and a reduced error rate. This proves that the proposed framework is robust to recognize the depression related content with different sizes of feature subsets. From the figures, it is observed that the proposed hybrid approach converges to the best solution (minimal error rate) for both the datasets. The convergence of other approaches is faster because of the premature convergence or early stagnation. Hence, the proposed FS approach tends to achieve high exploration and exploitation behavior in finding the optimal solution.

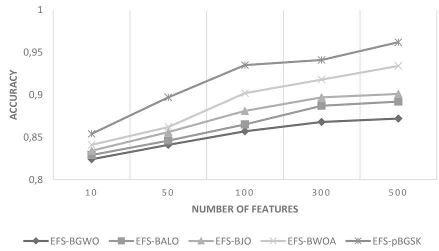
4.5.4. Algorithm complexity. The complexity of the proposed hybrid feature selection model is conditioned by two important factors: (i) time complexity and (ii) computational complexity. The time complexity of the model depends on the time consumed by the two core processes such as the EFS score calculation and optimization. The EFS score calculation is a one time process with running time complexity of $\mathcal{O}(wl)$, where w is the total number of features and l is the number of class labels. The optimization process with the BGSK algorithm depends on the time taken for initial population generation, updating knowledge vectors, etc. Therefore,

Table 5. Impact of pBGSK optimization on feature selection.

Dataset	Before optimization			After optimization		
	# features	Accuracy	F_1	# features	Accuracy	F_1
Sentiment 140	4958	0.853	0.847	**500**	**0.942**	**0.929**
SDD	6794	0.867	0.835	**500**	**0.959**	**0.952**

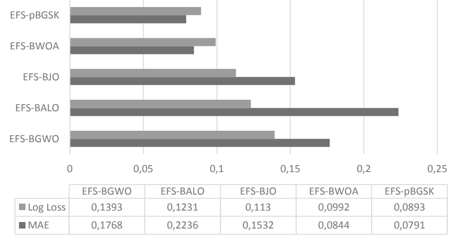


(a) NB classifier

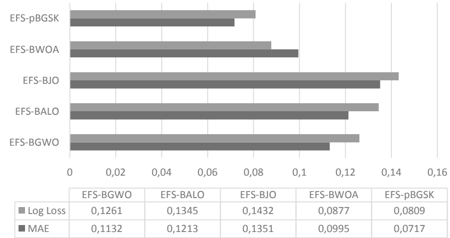


(b) SVM classifier

Fig. 4. Accuracy comparison for the SDD dataset.



(a) NB classifier



(b) SVM classifier

Fig. 5. MAE and log-loss comparison for the SDD dataset.

the time complexity of each iteration is $\mathcal{O}(w \times \mathrm{NP} + \mathrm{Cof} \times \mathrm{NP})$, where 'Cof' represents the cost of the fitness function. The computational complexity of the model typically considers the number of function evaluations (NFE) and the population count, and can be defined as $\mathcal{O}(w \times \mathrm{NFEs} + \mathrm{Cof} \times \mathrm{NFEs})$. This implies that the model quality is determined based on the population count, the cost of the fitness function and the number of function evaluations performed. Hence, while using pBGSK, the population count reduces gradually for each FE which in turn minimizes both the time and computational complexity over each iteration.

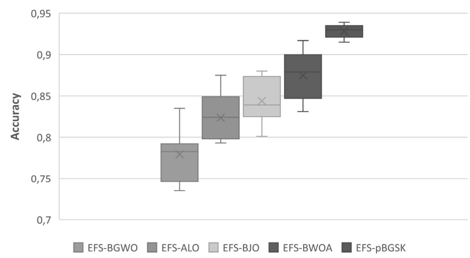
4.5.5. State-of-the art comparison. As part of the evaluation process, the proposed model is compared with an existing depression detection system (Tadesse *et al.*, 2019), where LDA topics, LIWC dictionary and N-gram features are employed. The system used various machine learning algorithms for text classification including a multi-layer perceptron (MLP), an SVM, etc. Table 6 presents the result of the comparison of the proposed unigram based hybrid feature selection model with the aforementioned system. The maximum performance scores obtained by Tadesse *et al.* (2019) are 0.91 (accuracy) and 0.93 (F1 score) using the combination of LIWC+LDA+bigram features with the MLP model. Similarly, using the SVM model, the system attained 0.90 (accuracy) and 0.91 (F1 score). For comparison, a simple MLP model with two hidden layers containing 4 and 16 neurons is constructed based on the proposed EFS-pBGSK model. The MLP algorithm also produces equivalently better performance, and it is clear that the proposed model can be able to detect depressive posts with better performance even using simple unigram features.

5. Conclusion

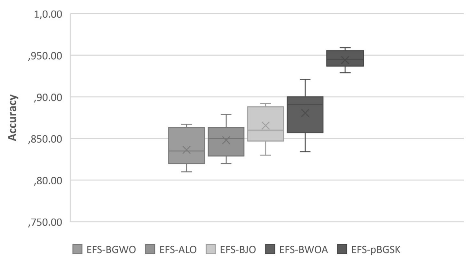
The depression classification using text sequences is a challenging task. In text classification, the most crucial part is to select the most significant feature subset from the sparse feature space. The proposed EFS-pBGSK approach focused on selecting an optimal feature subset for text depression classification by encapsulating the advantages of both filter and wrapper methods. The filter method has successfully ranked the features based on EFS scores. The pBGSK algorithm reduces the dimensionality of the feature space by selecting the minimal and optimal set of features for text classification. The performance of the proposed work was tested on two benchmark

Table 6. Comparison with a state-of-the-art approach.

	Dataset	Features	Model	Accuracy	F1 score
Tadesse *et al.*, 2019	Reddit posts	LIWC+LDA+bigram	SVM	0.900	0.910
			MLP	0.910	0.930
Proposed	Sentiment 140	Unigram	EFS+pBGSK+SVM	0.942	0.929
			EFS+pBGSK+MLP	0.937	0.921
	SDD	Unigram	EFS+pBGSK+SVM	0.962	0.952
			EFS+pBGSK+MLP	0.946	0.929

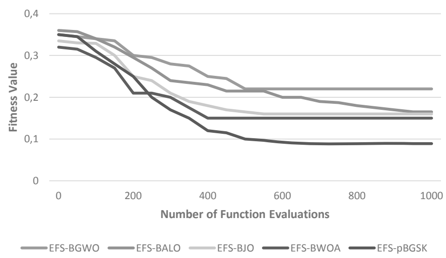


(a) Sentiment140

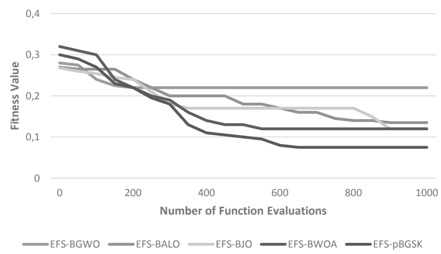


(b) SDD

Fig. 6. Accuracy box plots of various hybrid approaches.



(a) Sentiment140



(b) SDD

Fig. 7. Convergence curves of different hybrid feature selection approaches.

datasets for depression detection from Twitter and Reddit platforms. Finally, the experimental results show that the proposed model effectively reduces the number of features and increases the performance of depression classification in terms of accuracy and F_β scores. In the future, the depression can be detected using multimodal features (images, audio, video, etc.) instead of text only features in a more effective manner.

Acknowledgment

The authors are grateful to the anonymous reviewers for their constructive comments and suggestions, which helped us to improve the quality of the manuscript. The authors would like to extend their gratitude to the Management and Principal of Mepco Schlenk Engineering College (Autonomous), Sivakasi, for providing ample facilities and support to carry out this research work.

References

Agrawal, P., Ganesh, T. and Mohamed, A. (2021). A novel binary gaining-sharing knowledge-based optimization algorithm for feature selection, *Neural Computing and Applications* **33**: 5989–6008.

Asim, M., Wasim, M., Sajid Ali, M. and Rehman, A. (2017). Comparison of feature selection methods in text classification on highly skewed datasets, *2017 1st International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT), Karachi, Pakistan*, pp. 1–8.

Babu, N. and Kanaga, E. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: A review, *SN Computer Science* **3**: 74.

Burdisso, S., Errecalde, M. and Montes, M. (2019). A text classification framework for simple and effective early depression detection over social media streams, *Expert Systems with Applications* **133**: 182–197.

Chen, J., Huang, H., Tian, S. and Qu, Y. (2009). Feature selection for text classification with naïve Bayes, *Expert Systems with Applications* **36**(3): 5432–5435.

Chiong, R., Satia Budhi, G., Dhakal, S. and Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts, *Computers in Biology and Medicine* **135**: 104499.

Deng, X., Li, Y., Weng, J. and Zhang, J. (2019). Feature selection for text classification: A review, *Multimedia Tools and Applications* **78**: 3797–3816.

Derek, A. and David, M. (2020). Support vector machine, *in* A. Mechelli and S. Vieira (Eds), *Machine Learning*, Academic Press, Chicago, pp. 101–121.

Ding, Y., Chen, X., Fu, Q. and Zhong, S. (2020). A depression recognition method for college students using deep integrated support vector algorithm, *IEEE Access* **8**: 75616–75629.

Durgalakshmi, B. and Vijayakumar, V. (2020). Feature selection and classification using support vector machine and decision tree, *Computational Intelligence* **36**: 1480–1492.

Emary, E., Zawbaa, H. and Aboul Ella, H. (2016a). Binary ant lion approaches for feature selection, *Neurocomputing* **213**: 54–65.

Emary, E., Zawbaa, H.M. and Hassanien, A.E. (2016b). Binary grey wolf optimization approaches for feature selection, *Neurocomputing* **172**: 371–381.

Friedrich, M. (2017). Depression is the leading cause of disability around the world, *Journal of the American Medical Association (JAMA)* **15**: 1517.

Gao, Z., Xu, Y., Meng, F., Qi, F. and Lin, Z. (2014). Improved information gain-based feature selection for text categorization, *4th International Conference on Wireless Communication, VITAE, Aalborg, Denmark*, pp. 1–5.

Hayyolalam, V. and Kazem, A. (2020). Black widow optimization algorithm: A novel meta-heuristic approach for solving engineering optimization problems, *Engineering Applications of Artificial Intelligence* **87**: 103249.

Hussain, J., Satti, F., Afzal, M., Khan, W., Bilal, H., Ansaar, Z., Ahmad, H., Hur, T., Bang, J., Kim, J., Park, G., Seung, H. and Lee, S. (2019). Exploring the dominant features of social media for depression detection, *Journal of Information Science* **46**(6): 739–759.

Husseini Orabi, A., Buddhitha, P., Husseini Orabi, M.M. and Inkpen, D. (2018). Deep learning for depression detection of twitter users, *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, New Orleans, USA*, pp. 88–97.

Hussien, A.G., Oliva, D., Houssein, E.H., Juan, A.A. and Yu, X. (2020). Binary whale optimization algorithm for dimensionality reduction, *Mathematics* **8**(10): 1821.

Islam, M., Kabir, M., Ahmed, A., Kamal, A., Wang, H. and Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques, *Health Information Science and Systems* **6**(1): 8.

Kowal, M., Skobel, M. and Nowicki, N. (2018). The feature selection problem in computer-assisted cytology, *International Journal of Applied Mathematics and Computer Science* **28**(4): 759–770, DOI: 10.2478/amcs-2018-0058.

Li, B., Yan, Q., Xu, Z. and Wang, G. (2015). Weighted document frequency for feature selection in text classification, *International Conference on Asian Language Processing (IALP), Suzhou, China*, pp. 132–135.

Mohamed, A., Hadi, A. and Mohamed, A. (2020). Gaining-sharing knowledge based algorithm for solving optimization problems: A novel nature-inspired algorithm, *International Journal of Machine Learning and Cybernetics* **11**: 1501–1529.

Moorthy, U. and Gandhi, U. (2019). Forest optimization algorithm-based feature selection using classifier ensemble, *Computational Intelligence* **36**(4): 1445–1462.

Moradi, P. and Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy, *Applied Soft Computing* **43**: 117–130.

Parlak, B. and Uysal, A. (2021). A novel filter feature selection method for text classification: Extensive feature selector, *Journal of Information Science* **49**(1): 59–78.

Peng, H., Long, F. and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(8): 1226–1238.

Połap, D. and Woźniak, M. (2021). Red fox optimization algorithm, *Expert Systems with Applications* **166**: 114107.

Prachi, A., Abutarboush, H., Ganesh, T. and Mohamed, A. (2021). Metaheuristic algorithms on feature selection: A survey of one decade of research (2009–2019), *IEEE Access* **9**: 26766–26791.

Rajalakshmi, R. and Aravindan, C. (2018). A naive Bayes approach for URL classification with supervised feature selection and rejection framework: NB for URL classification with FS and RF, *Computational Intelligence* **34**(1): 363–396.

Rao, R. (2016). Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems, *International Journal of Industrial Engineering Computations* **7**: 19–34.

Rehman, A., Javed, K. and Babri, H. (2017). Feature selection based on a normalized difference measure for text classification, *Information Processing and Management* **53**(2): 473–489.

Sanasam, R., Murthy, H. and Gonsalves, T. (2010). Feature selection for text classification based on Gini coefficient of inequality, *Proceedings of Machine Learning Research* **10**: 76–85.

Shen, J. and Rudzicz, F. (2017). Detecting anxiety through Reddit, *Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality, Vancouver, Canada*, pp. 58–65.

Suthaharan, S. (2016). *Machine Learning Models and Algorithms for Big Data Classification*, Springer, Boston, chapter “Support vector machine”, pp. 207–235.

Tadesse, M., Lin, H., Xu, B. and Yang, L. (2019). Detection of depression-related posts in Reddit social media forum, *IEEE Access* **7**: 44883–44893.

Thirumoorthy, K. and Muneeswaran, K. (2020). Optimal feature subset selection using hybrid binary Jaya optimization algorithm for text classification, *Sādhanā* **45**(201).

Thorstad, R. and Wolff, P. (2019). Predicting future mental illness from social media: A big-data approach, *Behavior Research Methods* **51**: 1586–1600.

Trotzek, M., Koitka, S. and Friedrich, C. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, *IEEE Transactions on Knowledge and Data Engineering* **32**(3): 588–601.

Unler, A., Murat, A. and Chinnam, R. (2011). MR2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification, *Information Sciences* **181**(20): 4625–4641.

Uysal, A. (2018). On two-stage feature selection methods for text classification, *IEEE Access* **6**: 43233–43251.

Wang, W., Chen, X., Musial, J. and Blazewicz, J. (2020). Two meta-heuristic algorithms for scheduling on unrelated machines with the late work criterion, *International Journal of Applied Mathematics and Computer Science* **30**(3): 573–584, DOI: 10.34768/amcs-2020-0042.

William, D. and Suhartono, D. (2021). Text-based depression detection on social media posts: A systematic literature review, *Procedia Computer Science* **179**: 582–589.

Xue, B., Zhang, M., Browne, W. and Yao, X. (2016). A survey on evolutionary computation approaches to feature selection, *IEEE Transactions on Evolutionary Computation* **20**(4): 606–626.

Zhu, X., Wang, Y., Li, Y., Tan, Y., Wang, G. and Song, Q. (2019). A new unsupervised feature selection algorithm using similarity-based feature clustering, *Computational Intelligence* **35**(1): 2–22.

Santhosam Kavi Priya is an associate professor in the Department of Computer Science and Engineering at Mepco Schlenk Engineering College (Autonomous). She holds a PhD from Anna University, Chennai, India, an ME degree in computer science and engineering from Mepco Schlenk Engineering College (Autonomous), and a BE degree in computer science and engineering from MK University, Tamil Nadu, India. She has nearly 18 years of teaching experience. Her research interests include data mining, machine learning, the Internet of things and wireless sensor networks.

Kasirajan Pon Karthika, PWD, works at the Department of Computer Science and Engineering at Mepco Schlenk Engineering College (Autonomous), Tamil Nadu, India. She holds an ME degree in computer science and engineering and a BTech degree in information technology, both from Mepco Schlenk Engineering College (Autonomous), Tamil Nadu, India. Her research areas include data mining, text mining and machine learning.

Received: 6 March 2022
Revised: 23 May 2022
Accepted: 3 July 2022