I.J. Modern Education and Computer Science, **2023, 5, 60-73**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2023.05.05

A Novel Algorithm for Stacked Generalization Approach to Predict Neurological Disorder over Digital Footprints

Tejaswita Garg*
School of Studies in Computer Science & Applications, Jiwaji University-Gwalior, Madhya Pradesh, 474011, India
E-mail: tejaswitagarg@gmail.com
ORCID ID: https://orcid.org/ 0009-0005-5097-4326
*Corresponding Author

Sanjay K. Gupta
School of Studies in Computer Science & Applications, Jiwaji University-Gwalior, Madhya Pradesh, 474011, India
E-mail: sanjaygupta9170@gmail.com

Received: 09 May, 2023; Revised: 10 June, 2023; Accepted: 24 July, 2023; Published: 08 October, 2023

Abstract: Digital footprints track online behaviors of an individual when communicating over social media platforms. In this paper, sentiment classification is carried out over online posts and tweets to pre detect whether a person is having neurological disorder or not. This study proposed a Hybrid Optimized Model Ensemble STACKed (HOMESTACK) algorithm built on stacked generalization approach that uses stacking and blending ensemble learning technique. The model is then evaluated over two datasets (Reddit Dataset1 & Twitter Dataset2) that include varied number of tweets. The pre-processing of the data and feature extraction is carried out to get cleaned text and vector corpus. The proposed HOMESTACK algorithm is then applied over training data using four base classifiers as Support Vector, Random Forest, K-Nearest Neighbor and CatBoost along with a Meta classifier as Logistic Regression. The testing data is then fed to the tuned model to compare the classification results and analysis. Also, Stacking and Blending ensemble frameworks and algorithms are proposed in this study. Execution time and metric evaluation are calculated in respect of Accuracy, Precision, Recall and F1-score. The experimental results clearly show that the proposed HOMESTACK algorithm performed better over chosen datasets as compared to blending ensemble and standalone machine learning classifiers.

Index Terms: Digital Footprint, Stacked Generalization, Machine Learning, Word Embedding, Stacking and Blending Ensemble, Sentiment Analysis, Neurological Disorder.

1. Introduction

Humans are social creatures by nature. However, as compared to the earlier communication, modern era has undergone a significant transition. Social media is now vital medium for communication that is utilized by practically all facets of society [1]. The highly rated online networking platforms include YouTube, Reddit, Twitter, Instagram and Facebook. Reddit is a digital news accumulation, discussion, and content-rating web service based in the United States. The website's material, which includes links, text entries, photographs, and videos, is contributed by registered users and is then rated by other users. On other hand Twitter, a global conversation forum, has become an established social media data collection in the discipline of research. As a result, it makes it easier for users to interact with institutions or organizations. Because of the widespread usage of online communication platforms, people have the ability to provide comment about circumstances, occasions, things and facilities [2]. These comments are frequently relied on user's knowledge, which may include favorable or unfavorable perceptions of goods or service. Finding negative user feedback is essential to the development of the organizations and these suggestions will then aid businesses in betterment of their goods and services, enabling them to sell more. Therefore, it is critical to include user input gathered from websites and social media. Through text data, sentiment analysis may effectively disclose people' opinions (whether positives, negatives, or neutrality) about a good or service.

Depression is real ailment that affects a person's physical and mental feelings. Despite the fact that depression is a curable mental disorder, many individuals lack the chance to access care quickly for a wide range of reasons. The statue of the sickness is not identified at an early stage because the typical treatment demands close connection. However, more peoples are using social media, and they feel free to express their emotions on networking sites like Facebook, Twitter, Reddit, blogs, etc. Users are able to identify their bad emotions early on because of this [3]. However, in the majority of cases, early professional assistance can relieve mental symptoms (such as low self-esteem & worry) and treat somatic issues (such as digestion problems and sleeping disturbances).

Early detection of neurological disorder which seeks for evaluation and treatment can greatly increase the likelihood that signs and the underlying condition will be controlled, as well as lessen the negative effects on one's health as well as their personal and professional lives. However, it can be difficult and resource intensive to identify depression symptoms. The majority of current methods rely on clinical research and health surveys conducted by hospitals or medical groups, which use psychological or neurological evaluation to forecast the presence of mental disorders. Fig. 1 shows the basis steps carried out in sentiment analysis approach.

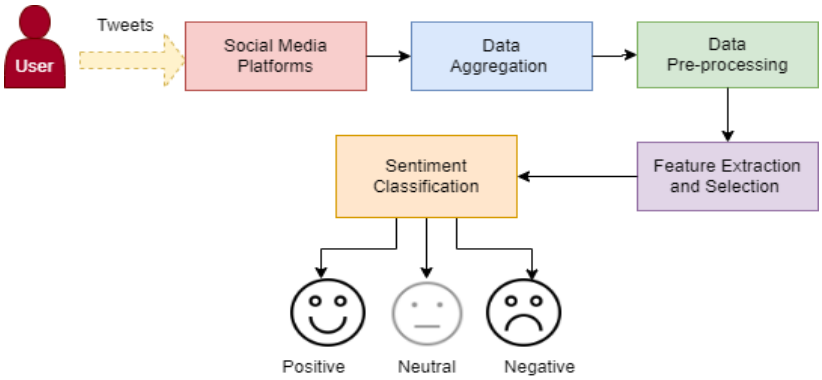


Fig. 1. Flowchart of Sentiment Analysis.

It is possible to think about sentiment analysis as a combined process that involves computational linguistics, natural language processing (NLP), and text mining [4]. Finding the text's underlying emotion, whether it's favorable, unfavorable, or neutral—helps. The use of both syntactic and semantic data, machine learning algorithms is applied for finding the sentiment through sentiment analysis approach. The classification process will employ training and testing dataset that evaluates the accuracy of classification model. The dataset's text must be turned into vectors in order to train the model; an old style BoW vectorizer is straight forward, but its output has huge dimensions and is sparse.

Word embeddings are presented to remove the challenges associated that come with addressing these vectors and feature sets. The distributional hypothesis states that the word correlation data are used to create distributed, fixed-length word vectors called word embeddings [5]. Here, Word2Vec, GloVe and FastText are the three different word embedding models that were employed, along with four classifiers are empirically compared in this study. A new stacking and blending ensemble learning approach with improved accuracy than standalone models is also suggested. If the algorithms are chosen to generate the hybrid stack based ensemble that yields good accuracy scores. By utilizing several base classifiers and Meta learners, we are creating stack-based ensemble classifiers in the current research as well.

The following is a summary of the study's major contributions:

- This work investigates the feasibility of employing pre-processing over lexical analysis and assesses the effectiveness of a stacked and blending ensemble for the categorization of posts or tweets using sentiment analysis.
- For the purpose of classifying sentiment on two datasets (Reddit & Twitter posts), a novel Hybrid Optimized Model Ensemble STACKed (HOMESTACK) algorithm of machine learning models including SVC, RFC, KNN, CBC, and LR is proposed as Algorithm 1.
- With binary classification tasks, LR performs best. Hence, the proposed stacked ensemble makes use of LR capabilities as a Meta learner to effectively combine the predictions provided by four base learners.
- We also assess how well our suggested strategy performs in comparison to related studies on the pre-determination of such neurological disorders over social media platforms.

The objective is to analyze the extent and character of studies on neurological disorders like depression that use Reddit & Twitter platform as their main data source. To identify depression, the multiple researchers used a single machine-learning technique and paid attention to accuracy value in order to decide which strategy is optimal for the situation. The researchers also used a single dataset that was the same size for each approach. To improve classification accuracy, we use stacking based and blending ensemble algorithms in the current research to classify the social media postings sentiments. To increase the model's accuracy even more, stratified 10- fold validation is used with proposed

hybrid optimized model of stack-based ensemble classifiers, with SVC, RFC, KNN, and CBC serving as the four base classifiers and then Meta classifier as LR.

The following research concerns are addressed in this work:

1. How much data has been used to study depression and anxiety?
2. What text-based technique is most effective for spotting depression in its early stages?
3. How single algorithm behaves for smaller and larger dataset collections?

The organization of this study is as follows: The introduction for the technique behind sentiment analysis is discussed in Section 1. In next Section 2, existing approaches are discussed. In Section 3, the materials and techniques are briefly explained. The experimental procedure as proposed methodology includes data sources, text pre-processing operations and chosen machine learning based classifiers utilized for the sentiment analysis of Reddit & Twitter postings along with proposed algorithm are described in Section 4. The findings and discussions are covered in Section 5. Furthermore, Section 6 wraps up the conclusion.

2. Background Study

This section highlights the existing methods and approaches related to this research that is comparable with the proposed study. Although algorithms like machine or deep learning are widely used for analyzing the sentiment of tweets, further increasing efficiency is constantly a research priority. The use of ensemble classifiers is to increase their classification efficiency has drawn the interest of numerous researchers, and significant work has been done in this field.

On three separate data sets, four different ensemble strategies were used by this study [6] as bagging, voting, boosting and ensemble of stack based classifiers. The implementation of the stack-based ensemble model makes use of the base classifiers SVM, NB, KNN, and C4.5 as well as LR as meta-model. The results demonstrate that the proposed model performed better over other classifiers in terms of efficiency.

For the purpose of classifying text sentiment, author [7] likewise utilized various ensemble algorithms. Two levels of ensemble performance are used, one for feature selection and the other for classifiers. Two strategies are utilized at the features extraction level to minimize error and for effective feature selection because a poor feature selection can result in a poor classification. To improve classification accuracy, two algorithms are combined into an ensemble. Comparing the method to other machine-learning classifiers, it produces good results.

A KNN and NB ensemble was employed for sentiment classification by author [8]. By applying a KNN, NB, and SVM-based ensemble, results are further enhanced to a 95% accuracy level.

Seven base classifiers were used in the implementation of ensemble-based classifiers Bagging and AdaBoost in this study [9]. Classifier ensembles are frequently constructed using majority voting based ensembles.

As a result, a study of [10] proposed a hybrid sentiment analysis model that makes use of deep learning models and word embedding techniques. Here, the proposed hybrid model of CNN + BiLSTM, which was provided as an input to the authors' combination of FastText embedding and character embedding outperforms with accuracy of 82.14%.

Here, the dataset of Dutch citizen is used to identify depression through the machine learning model XGBoost. Only 5% of the dataset's instances are depressed, with the remainder being non-depressed cases. In this study, the numerous samples has been created using ROSE sampling approaches, over-O, under-U, over-under OU, and under-under UU sampling due to the imbalance of the dataset. Implementing the XGBoost model reveals that the Over & OU sampling approaches improve the diagnostic accuracy [11].

In this study [12], the scientists applied various machine learning models to identify depression in social media posts. The models LR, NB, RF, and SVM are employed. It may be concluded from the results that LR performs better.

In this study [13], In order to develop a predictive system to recognise melancholy based on tweets, the emotions of Arabic text is extracted. The performance of four classifiers, including NB, RF, AdaBoostM1, and Lib-linear, were evaluated; the latter produced outcomes with a better accuracy of 87.5%.

The authors of [14], the proposed work used word embedding with various classifier models to increase the effectiveness of sentimental analysis. In this work, the accuracy of the word2vec embedding model using a random forest classifier is 81%.

In order to identify the polarity of writing, sentiment classification on the Amazon data set is carried out in this study [15]. Through experimentation, it has been discovered that the hybrid classifier, which combines SVM with Random Forest, produces better results than either pure SVM or RF.

Although results from ensemble classification have been shown to be superior to those from standalone machine learning classifiers, accuracy of ensemble model depends heavily on the choice of standalone model. After doing a survey, it has discovered that there isn't much research being done on stack-based Ensemble classifiers. This paper also identifies worked with Reddit dataset is not that much explored till now and hence, it is necessary to look out disorder over such platforms as well.

In order to improve classification accuracy, the current paper uses a hybrid model that combines base learners with Meta learner in a stack-based ensemble technique. In order to improve accuracy scores, this paper also compares the results of blending and stacking all of the selected classifiers.

3. Materials and Methods

In this section, theoretical underpinnings of the text classification and sentiment analysis methodologies are then briefly discussed. Hence, the adopted methodology for word embeddings and classifiers are discussed below.

3.1. Word Embeddings

Word representations that are widely employed to improve the performance of NLP tasks benefit from word embeddings. The entirety of a word's useful syntactic and semantic properties is retained. It calculates the word similarity scores using cosine similarity and the Euclidean distance between two words. The cosine similarity scale goes from – 1 to 1 (opposite) as shown in equation (1). Through a model, two vectors that are highly similar are likely to produce the same outcomes.

$$similarity_{cos} = \cos(\theta) = \frac{\mathrm{A.B}}{||\mathrm{A}||.||\mathrm{B}||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

This section then briefly explains some of the word embedding-based predictors for various machine learning models [5].

3.1.1 Word2Vec

Word2Vec can use any of the two continuous skip-gram or continuous BoW (CBoW) model architectures. The model suggests the current word in the CBOW framework using a window of nearby context words. In the continuous skip-gram structure, the model forecasts the context words that will be in the immediate vicinity of the current word. Its goal is to create distributed illustrations that demonstrate some word reliance on other words. It is a two-layered network that decodes text and visualizes terms as vector. It takes input text-based data and outputs featured vector set. The feature vectors in this collection of vectors stand in for the words in the dataset. The CBoW and skip gram's objective function are defined as in equation (2) and (3) below:

CBoW Objective function is defined as:

$$J = -\log \hat{P}\ (\omega_t \mid \omega_{t-n+1} \dots\dots \omega_{t-1}) \quad (2)$$

Skip-gram's objective function is defined as:

$$J = \frac{1}{T} \sum_{t=1}^{T} \sum_{-n \le j \le n} \log p\ (\omega_{t+j} \mid \omega_t) \quad (3)$$

3.1.2 GloVe

GloVe, an unsupervised machine learning approach that extracts vectors from text, is another popular word embedding method. In accordance with the coexisting statistics of the terms in the dataset, Glove represents the words. The training is carried out via corpus-based universal word co-occurrence figures, and the results display some of outstanding linear substructures of the word vector space. The major flaw with these two embeddings is the generation of random vector inside a term rather than within the dataset.

Let i and j be two words and we need to find ratio of probabilities with some word k. So, the probability of co-occurrence between i and k words be P_{ik} over P_{jk}. It is described as in equation (4) to (7) below,

$$F\ (\omega_i, \omega_j, \widetilde{\omega_k}) = \frac{P_{ik}}{P_{jk}} \quad (4)$$

and, it can be simplified as,

$$\omega_i^T\ \widetilde{\omega_k} + b_i + \widetilde{b_k} = \log X_{ik} \quad (5)$$

The function, that is a least squares issue, is what it seeks to optimize:

$$J = \sum_{i,j=1}^{V} f\ (X_{ij})\ (\omega_i^T\ \widetilde{\omega_j} + b_i + \widetilde{b_j} - \log X_{ij})^2 \quad (6)$$

With

$$f\ (x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^{\alpha} & if\ x < x_{\max} \\ 1 & otherwise \end{cases} \quad (7)$$

ω_i word vector; X co-occurrence matrix; b bias terms; f weighting function; and $\tilde{\omega}$ context vector.

3.1.3 Fasttext

Fasttext, a development of Word2Vec, can, however get over this drawback. It generates a word using the n-gram technique. As a result, incorporating a term that is absent from the corpus is enhanced. This method's higher memory while operating is its key drawback. It embeds the words differently than GloVe by considering each as a group of letter n-grams instead of a single word. It can learn new words that are both unusual and unfamiliar thanks to this trait. Indeed, with SkipGram, word representations through it with a scoring function s parametrize the likelihood of the contextual, giving a word t:

$$s\left(\omega_t, \omega_c\right) = u^T \omega_t . v_{\omega_c} \tag{8}$$

using the matrices such as input embedding as u and output as v. The scoring function in FastText is now:

$$s\left(\omega_t, \omega_c\right) = = \sum_{g \in \mathcal{G}\, \omega_t} z_g^T \; . v_{\omega_c} \tag{9}$$

with z_g as the vectors of the g-th n-gram and $\mathcal{G}_{\omega_t}$ as the collection of n-grams in term ω_t. The vectors of word context ω_c is v_{ω_c} .

Hence, the dataset in this study was represented using pre-trained three word embedding techniques.

3.2. Classifiers

Datasets are typically split into two categories as training and testing datasets. The classifier creates a model using the training dataset which is then evaluated against by the test dataset to make sure the trained model has the necessary accuracy. This model can be utilized to estimate the label with actual data after achieving the necessary accuracy. The machine learning algorithms employed in this study are listed below.

3.2.1 Base Classifier

Support Vector Machine Classifier (SVC) selects the extreme vectors and points that aid in the creation of the hyper plane. Support vectors are used in situations like in this study [16]. *Random Forest Classifier (RFC)* depends on the majority votes of the forecasts, it predicts the outcome by using the predictions from each tree[17].

K-Nearest Neighbor Classifier (KNN) states that all of the data is stored via the K-Nearest Neighbor algorithm, which classifies fresh data points based on similarities [18].

CatBoost Classifier (CBC) can automatically transform categorical data into numerical vectors without any pre-processing. When operating with dataset that contain a large number of categorical variables, this can save a tonne of time and effort [19].

3.2.2 Ensemble Classifiers

To enhance generalisation and predictions, ensemble learning integrates the results of individual models. The three important facets of learning—statistical, computational, and representable are enhanced by ensemble learning [20, 21]. As biased data is trained on single model is less likely to be used when using ensemble approaches, which lessen the threat of information inaccurate portrayal by merging numerous models. Unlike ensemble methods, which may carry out searches using random seeds and different starting positions with fewer system resources, most learning techniques look locally for an answer, which restricts the best response. To create a better classifier, we combine various base classifiers.

Meta Classifier: Logistic Regression Classifier

Binary logistic regression, which is how LR operates, is used to forecast the likely outcome of the target variable. There are only two feasible classes because of the dualistic nature of the dependent variable. The dependent variable is, to put it simply, a Boolean value, having values stored either as 1 (that signifies success/yes) or 0 (that also signifies failure/no). In the proposed work, we have used Ensemble classifiers with base classifiers SVC, RFC, KNN and CBC whereas LR as Meta classifier.

Stacking

The meta-model in the stacking ensemble, also known as "Stacking," used here as HOMESTACK as proposed model that is trained using CV on the underlying models' out-of-fold predictions. The learning process in detail is shown in Fig. 2. The training set is first divided into k folds; in this study, a stratified 10-fold CV is used. The left over fold is used to make predictions after each base model has been fitted using (k-1) folds. Now, repeating that procedure results in all k folds. The meta-model is then trained using the estimates from all base models and the results from the training set.

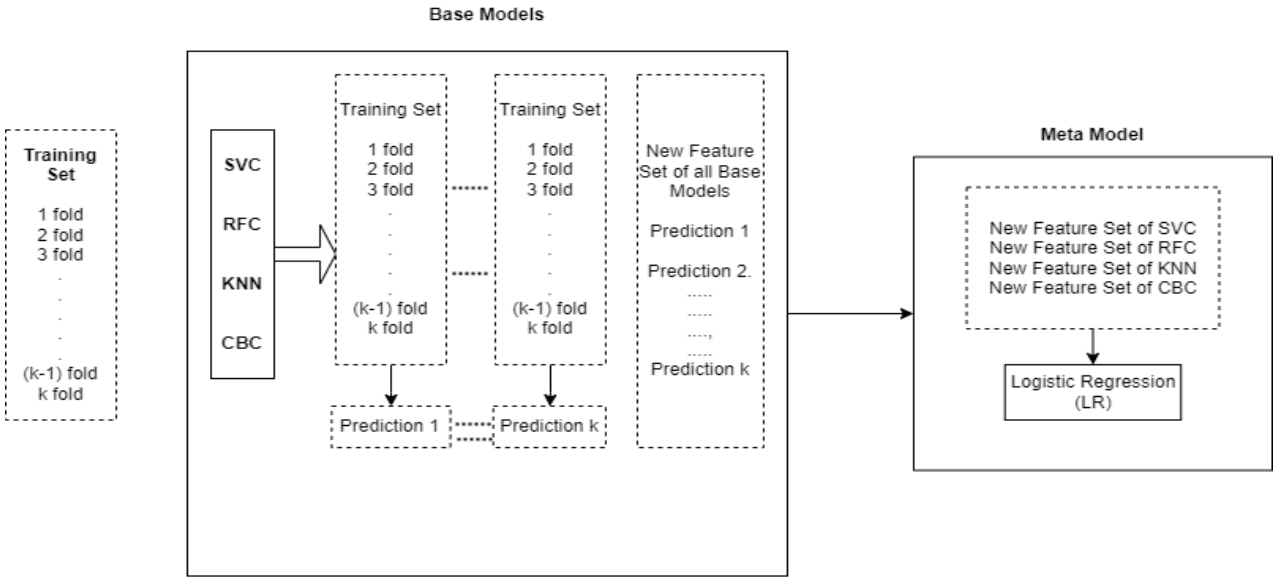


Fig. 2. A framework of Stacking ensemble learning approach with k-fold validation for proposed model.

Blending

The meta-model is fitting here on estimates on such a holdout testing dataset in the blending ensemble, also known as Blending. The training set is divided into two sections, as seen in fig. 3, one of which serves as the testing set and the other of which is utilized to train basic classifiers. The results from the testing set are combined with the predictions produced by the basic models on the training dataset as input features to fit the meta-model.

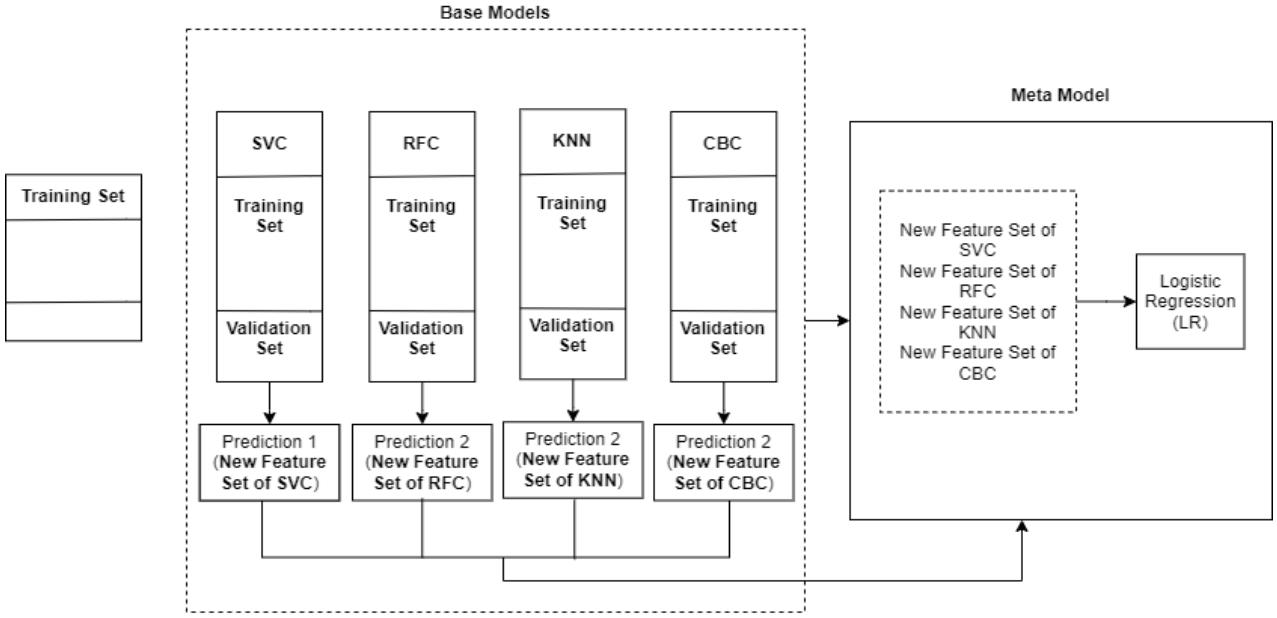


Fig. 3. A framework of Blending Ensemble learning approach for the proposed model.

Typically, it is employed to improve the accuracy and efficacy of fundamental machine learning classifiers. The objective function is rarely represented by a single hypothesis, but ensembles of several hypotheses can more closely resemble the target function.

4. Proposed Methodology

This section discusses the suggested technique; it includes pre-processing with the help of machine learning based word embedding; followed by modeling approaches and then proposed algorithm applied on both the datasets. In order to attain higher performance impacts, this research builds a hybrid stacked ensemble model HOMESTACK with stratified 10-fold validation approach and blending classifier model appropriate for recognizing depressed people in social media. The proposed architecture of the model is discussed as shown in Fig. 4 below.

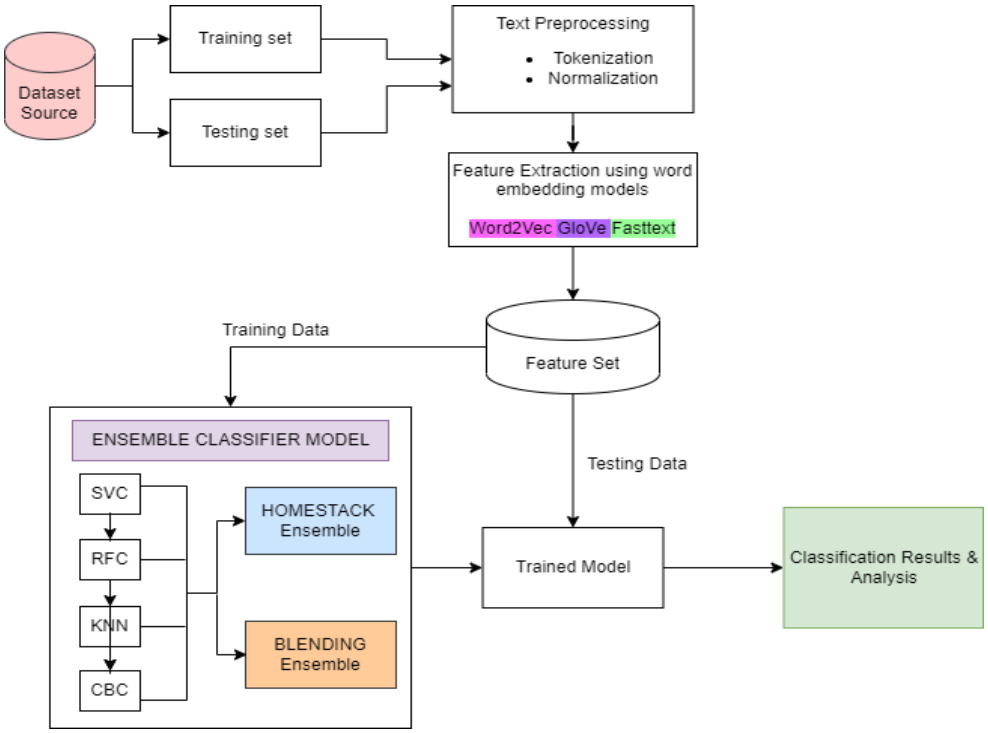
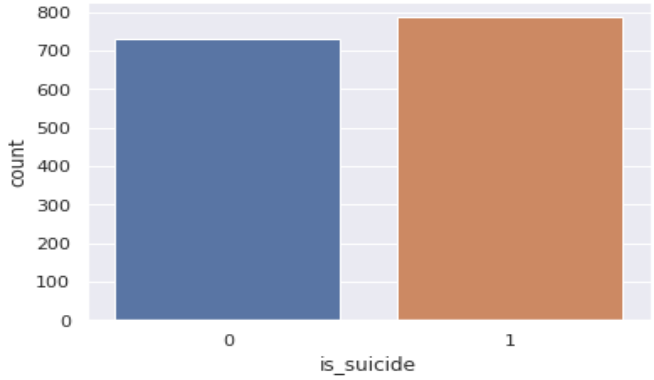


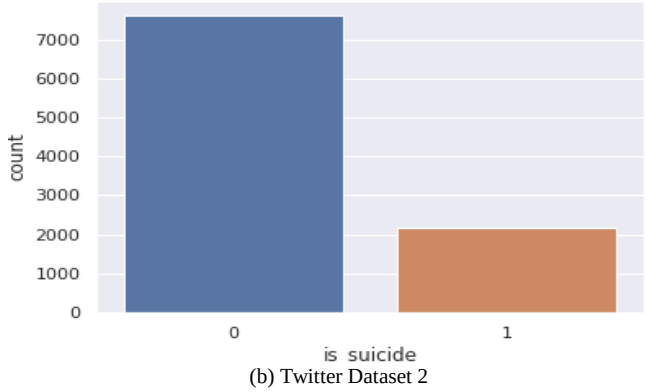
Fig. 4. Proposed architecture using Stacked Generalization approach for the proposed model.

4.1. Data Collection

In this study for accessing the dataset, social media platform as Dataset 1 (Reddit posts) are collected from the publicly accessible Github dataset that were depressive and non-depressive in nature were collected. There are 1516 posts in the dataset as a whole shown in Fig. 5(a) whereas Dataset 2(Twitter posts) are collected from the data source Kaggle, an open repository that includes 10308 Twitter sampled posts included in this study as shown in Fig. 5 (b) below.



(a) Reddit Dataset 1



(b) Twitter Dataset 2

Fig. 5. Dataset classification with two class labels as is_suicide disorder or non-disorder for (a.) Reddit dataset 1 and (b.) Twitter dataset 2.

4.2. Text Preprocessing

Preprocessing is required to ensure precision because the dataset is noisy. The dataset underwent preprocessing by having html links and mentions (#, @, https,: etc.), numerals, stop word, and punctuation removed. It has performed word separation and word extraction techniques such as tokenization, lemmatization, and stemming to extract valid terms from the tweets. The experimental results after cleaning the text are shown in Table 1.

Table 1. Tokenized text generation after pre-processing for 1 (a). Reddit dataset and 1(b) Twitter dataset.

a).

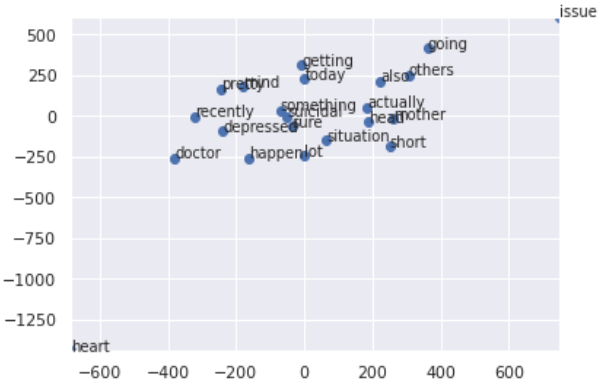
	selftext_clean	**is_suicide**	**tokenized_text**
1510	think spiritual person also think ever belongi…	1	[think, spiritual, person, also, think, ever, …..
1511	Every night guy want ice cream dinner everyone..	1	[every, night, guy, want, ice, cream, dinner, ….
1512	Would like say wa shook knew ha habit consumin..	1	[would, like, say, wa, shook, knew, ha, habit….
1513	Take anymore wanting buy pocket pistol similar ..	0	[take, anymore, wanting, buy, pocket, pistol, …
1515	Feel like people controlling every aspect life …	1	[feel, like, people, controlling, every, aspect ….

b.)

	clean_message	**is_suicide**	**tokenized_text**
8282	cannot diagnose distance medical professional ….	1	[cannot , diagnose, distance, medical, professional ..
3343	watch movie night whoaaaaaa …	0	[watch, movie, night, whoaaaaa ….
8791	honestly hit depression week someone cuddle …	1	[honestly, hit, depression, week, someone, cuddle...
2244	every song heidi montag ever release …	0	[every, song, hiedi, montag, ever, release , …
1559	blowdart say artistic persuasion autistic persuasion	0	[blowdart, say, artistic, persuasion, autistic, ….

4.3. Feature Extraction

Instead of only utilizing vectorizer, word embeddings are adopted in this work for achieving increase accuracy. Word2Vec, GloVe and Fasttext are machine learning based embedding techniques that are employed in the proposed work through python using t-SNE plots as shown in Figure 6-7. The semantic meaning of terms is reflected in real number vectors via embedding and similarly, the representation of words will indeed be close to one another.

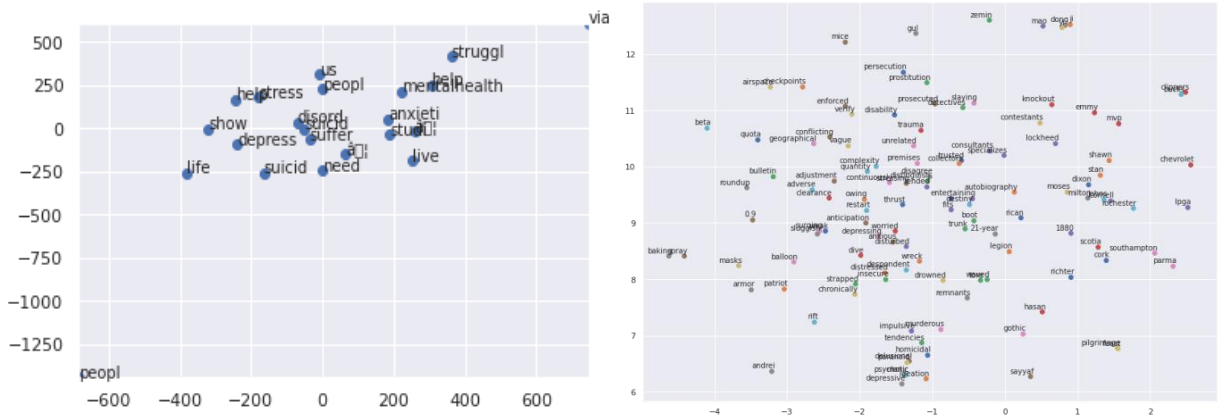


(a) Word2Vec

(b) GloVe

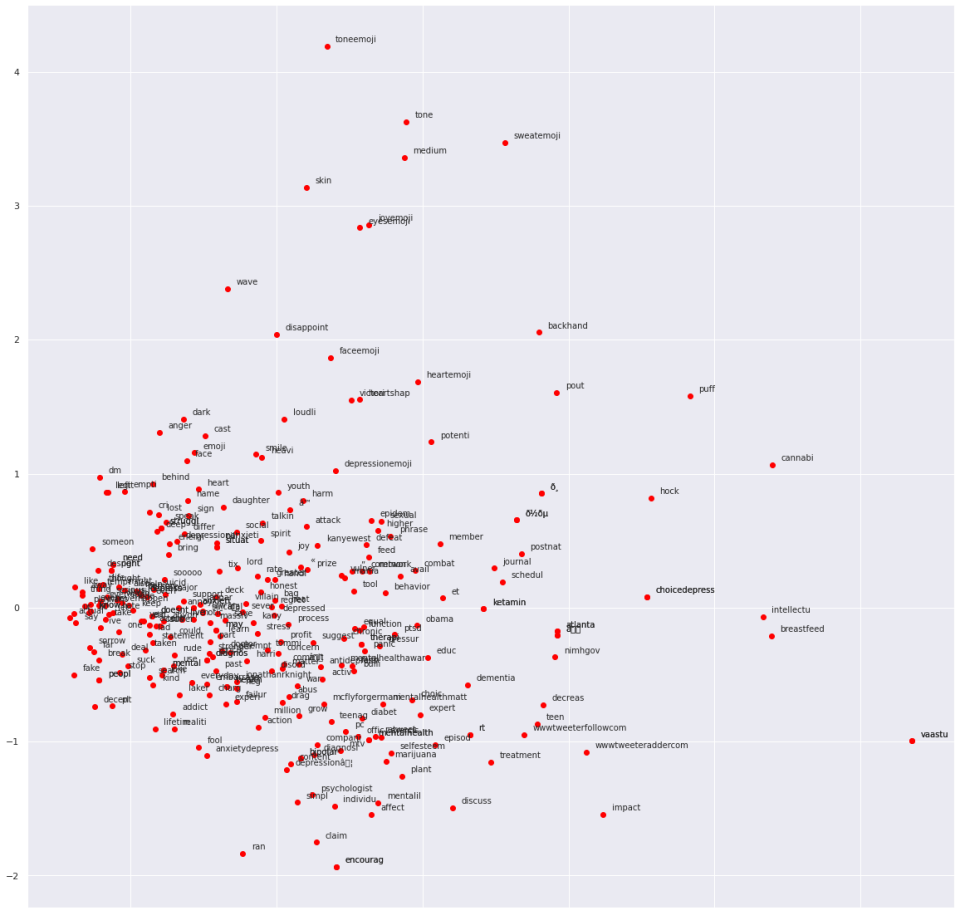
(c) Fasttext

Fig. 6. Displaying closed words for Reddit dataset 1 using word embeddings as (a) Word2Vec, (b.) GloVe and (c.) Fasttext.



(a) Word2Vec

(b) GloVe



(c) Fasttext

Fig. 7. Displaying closed words for Twitter dataset 2 using word embeddings as (a.) Word2Vec, (b.) GloVe and (c.) Fasttext.

4.4. Classification

Five classifiers were used for the binary classification problem: SVC, RFC, KNN and CBC as base classifiers and LR as a Meta model Classifier. Researchers are looking for the ideal stacking of embedding model for these classifiers as prior work for detecting neurological pre symptoms from posts yielded superior results with these algorithms.

4.5. Proposed Ensemble Classifiers

Algorithm 1 and Algorithm 2 explains the strategy used to implement the model. By removing URLs, digits, punctuation, Re-tweets, special characters, and hashtags, tweets are cleaned up and pre-processed. After cleaned text, techniques for Feature Extraction using three word embedding models as Word2Vec, GloVe and Fasttext are performed that expresses a feature as an array A of uni-grams. It generates aggregated sentence vectors based on word vectors for each word in a sentence. Each sentence has different number of array vectors which may cause an error while we train the model. Compute sentence vectors by averaging the word vectors for the words contained in the sentence to get updated feature set that may significantly boost ensemble classifier and machine learning efficiency with minimal overhead. All of the models are implemented over the chosen data set.

The proposed Hybrid Optimized Ensemble Stacked Algorithm “HOMESTACK” and Blending Algorithm are implemented using python libraries sklearn and modules. The effectiveness of the models is represented by the calculation of performance metric scores.

Hybrid Optimized Model Ensemble Stacked (HOMESTACK) Algorithm 1

Input: Reddit posts with Positive or Negative labels.
Output: Accuracy, Precision, Recall and F1-Score of Classified Posts.
Step1. Collection of Reddit Posts from Suicide and Depression Dataset from Github dataset source. Split data into training and testing sets.
Step2. Tokenization and Pre-processing of Posts.
Step3. Generate aggregated sentence vectors based on word vectors for each word in a sentence using three word embedding models as Word2vec, GloVe and Fasttext that expresses a feature as an array **A** of uni-grams.
Step4. Each sentence has different number of array vectors which may cause an error while we train the model. Compute sentence vectors by averaging the word vectors for the words contained in the sentence as **Anew**.
Step5. Create HOMESList S1 to build stacked ensemble learning model:
S1-> HOMESList (SVC, RFC, KNN, CBC)
Step6. For **Anew** updated feature set with LR Meta classifier
Sort posts using the **HOMESTACK** stacked ensemble approach with S1.
Step7. Tuning the Meta Classifier.
Step8. Sort posts into categories using unique machine learning classifiers SVC, RFC, KNN and CBC.
Step9. Examine and contrast the performance metric score of **HOMESTACK algorithm** and individual classifiers.
Step10. END

Blending Algorithm 2

Input: Reddit posts with Positive or Negative labels.
Output: Metric Evaluation Score of Classified Posts.
Step1. Collection of Reddit Posts from Suicide and Depression Dataset from Github dataset source. Split data into training and testing sets.
Step2. Tokenization and Pre-processing of Posts.
Step3. Generate aggregated sentence vectors based on word vectors for each word in a sentence using three word embedding models as Word2vec, GloVe and Fasttext that expresses a feature as an array **A** of uni-grams.
Step4. Each sentence has different number of array vectors which may cause an error while we train the model. Compute sentence vectors by averaging the word vectors for the words contained in the sentence as **Anew**.
Step5. Create and build blending ensemble learning model:
S1-> BlendList(SVC, RFC, KNN, CBC)
Step6. Fit, train and make predictions with the blending ensemble.
Step7. For **Anew** updated feature set with LR meta classifier
Sort posts using the **Blending** ensemble approach with S1.
Step8. Sort posts into categories using unique machine learning classifiers.
Step9. Examine and contrast the performance metric score of **blending algorithm** with stacking classifiers.
Step10. END

5. Results and Discussions

The Hybrid Optimized Model Ensemble STACKed (HOMESTACK) algorithm represented in Algorithm 1 is implemented on both datasets as Suicide and depression Reddit Dataset1 used from github source and on Sentimental Analysis for Tweets on Twitter Dataset2 used from Kaggle source. Utilizing LR Meta classifier on the selected dataset, a hybrid model composed of a classifier and stack-based ensembles of SVC, RFC, KNN, and CBC base classifiers are created. For all ensemble methods with blending and stacking, classification performance and metric calculation are listed in the Table 2 and 3, respectively.

Table 2. Comparative analysis of accuracy scores of proposed stacked model with standalone base classifiers.

		Word2Vec	**GloVe**	**Fasttext**
SVC	Dataset 1	0.604306	0.767756	0.743267
	Dataset 2	0.985858	0.994353	0.984003
RFC	Dataset1	0.943952	0.999505	0.995133
	Dataset2	0.976283	0.989744	0.972458
KNN	Dataset1	0.736878	0.770341	0.777169
	Dataset2	0.970264	0.979458	0.949112
CBC	Dataset1	0.988598	0.999830	0.999842
	Dataset2	0.981992	0.991727	0.982773
Blending	Dataset1	0.749239	0.833652	0.778191
	Dataset2	0.945376	0.974160	0.918605
Proposed Stacked	**Dataset1**	**0.984029**	**0.999819**	**0.999384**
	Dataset2	**0.986700**	**0.995513**	**0.984035**

Table 3. Metric calculation of the proposed model for both datasets over three embeddings on two classes as non-disorder(0) or is_disorder(1).

			Precision	Recall	F1-Score
Reddit Dataset 1	Word2Vec	Class0	0.887	0.989	0.935
		Class1	0.986	0.864	0.921
	GloVe	Class0	0.989	1.0	0.994
		Class1	1.0	0.989	0.994
	Fasttext	Class0	0.955	0.927	0.954
		Class1	0.926	0.984	0.954
Twitter Dataset 2	Word2Vec	Class0	1.0	0.41	0.58
		Class1	0.84	1.0	0.91
	GloVe	Class0	0.0	0.0	0.0
		Class1	0.75	1.0	0.86
	Fasttext	Class0	0.0	0.0	0.0
		Class1	0.75	1.0	0.86

In the proposed work, it is stated on comparative analysis that the proposed HOMESTACK model with an Ensemble of SVC+RFC+KNN+CBC classifier responded more accurately having high scores with the chosen three embeddings. Table 2 gives comparison based on accuracy scores of individual and ensemble machine learning algorithm for both datasets over three different word embedding model. It is noticed that the CBC model is giving higher accuracy with small dataset of Reddit posts but it failed to perform well with enormous set of data for Twitter Dataset. On contrary, SVC model performed well with large dataset of Twitter posts but failed to give high performance for smaller set of data of Reddit posts.

Hence, we can conclude that amongst all, the proposed HOMESTACK model outperforms over two different types of datasets using three word embedding models and ensemble classifiers with higher accuracy than stand-alone base classifiers. Table 3 shows posts/tweets can be correctly classified and hence, gives metric calculations in terms of Precision, Recall and F1-score for each dataset using three embedding models over each class as non-disorder (0) and disorder (1).

The comparative analysis of proposed model with existing literature is discussed below in Table 4.

Table 4. Comparative Analysis of proposed work with existing literature.

Year	Author	Dataset	Approach	Results
2016	C. Troussas et al.	Twitter	Ensemble Learning	87%
2018	Y. Emre Isik et al.	Twitter	Stacked Ensemble	79.1%
2019	M. Naz et al.	Amazon	Ensemble with Forest Optimization	95%
2015	J. Prusa et al.	Twitter	Bagging and Boosting	82%
2020	Salur, M.U. and Aydin	Twitter	CNN & Bi-LSTM	82.14%
2020	A. Sharma and W. J. M. I. Verbeke	Dutch Citizen	XGBoost with Sampling methods	90%
2020	G. Geetha et al.	Twitter	Machine Learning	-
2019	S. Almouzini et al.	Arabic texts	RF, NB, Adaboost, Liblinear	87.5%
2018	O.B. Deho et al.	Tweets collected from U.S. Military base in Ghana	Random Forest	81%
2018	Y. Al Amrani et al.	Amazon	Random Forest Support Vector Machine (RFSVM)	83.4%
2023	**Proposed**	**Reddit & Twitter**	**Stacking and Blending Ensemble**	**99.8%**

As discussed in Table 4, this study shows that proposed model behaved well over varied nature of datasets using stacking ensemble architecture with an improved accuracy of 99.8% as compared to existing literature.

While standalone machine learning algorithms still outperformed ensemble classifiers in terms of the performance, stacking all individual models increased the accuracy of ensemble classifiers across the dataset. It is also noticed that Blending ensemble is also less effective than some base classifiers who performed well instead. The results clearly show that proposed HOMESTACK algorithm performed better as compared to blending of all the classifiers and single machine learning classifiers.

6. Conclusion

This paper aimed to evaluate the earlier identification phase of neurological disorder as depression using numerous machine learning algorithms. For this, the performance of emotion detection based on sentiment analysis is improved by the usage of stack-based ensemble learning classifiers. In the proposed study, a hybrid optimized ensemble stacked model is constructed using base SVC, RFC, KNN, CatBoost and Meta LR classifier. Furthermore, the blending ensemble learning is included in the feature set as a further step, and proposed HOMESTACK model is then applied with stratified 10-fold validation to feature set for achieving improved metric calculations. However, execution time with proposed stacking model is larger over all the base models but it has also proven that for the chosen two different natures of datasets, the resultant HOMESTACK algorithm performed well with increased accuracy score over three

word embeddings as Word2Vec, GloVe and Fasttext based on machine learning approach. Hence, this study concludes that the proposed algorithm behaved well for both the datasets. As indicated in the results sections, we are able to improve classification accuracy, and the results we get may be safely displayed since we can guarantee the equality of the facts detected. The outcomes should only be used for research since this issue doesn't constitute a clinical trial. In future, this work can be expanded with image and speech sentiment analysis with different embedding approaches and validation set with the proposed model.

References

[1] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin., 2019. A survey of sentiment analysis in social media. Knowl. Inf. Syst., vol. 60, no. 2, pp. 617-663.

[2] F. Hemmatian and M. K. Sohrabi., 2019. A survey on classification techniques for opinion mining and sentiment analysis. Artif. Intell. Rev., vol. 52, no. 3, pp. 1495-1545.

[3] NA, Reseena Mol, and S. Veni., 2022. A STACKED ENSEMBLE TECHNIQUE WITH GLOVE EMBEDDING MODEL FOR DEPRESSION DETECTION FROM TWEETS. Indian Journal of Computer Science and Engineering, Vol. 13 No. 2, e-ISSN : 0976-5166, p-ISSN : 2231-3850, DOI : 10.21817/indjcse/2022/v13i2/221302088

[4] A. Tariyal, S. Goyal, and N. Tantububay., 2018. Sentiment Analysis of Tweets Using Various Machine Learning Techniques. Int. Conf. Adv. Comput. Telecommun. ICACAT 2018, pp. 2–4, 2018, doi: 10.1109/ICACAT.2018.8933612.

[5] F. Almeida and G. Xexéo., 2019. Word embeddings: A survey. arXiv, no. 1991, 2019.

[6] C. Troussas, A. Krouska and M. Virvou., 2016. Evaluation of ensemble-based sentiment classifiers for Twitter data, in 7th International Conference on Information, Intelligence, Systems & Applications (IISA), Chalkidiki, pp. 1-6.

[7] Y. Emre Isik, Y. Görmez, O. Kaynar And Z. Aydin., 208. NSEM: Novel Stacked Ensemble Method for Sentiment Analysis, 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, pp. 1-4.

[8] M. Naz, K. Zafar, A. Khan., 2019. Ensemble Based Classification of Sentiments Using Forest Optimization Algorithm, vol. 4, no. 2, pp. 1-13. https://doi.org/10.3390/data4020076

[9] J. Prusa, T. M. Khoshgoftaar and D. J. Dittman., 2015. Using Ensemble Learners to Improve Classifier Performance on Tweet Sentiment Data. IEEE International Conference on Information Reuse and Integration, San Francisco, pp. 252-257.

[10] Salur, M.U. and Aydin, I., 2020. A novel hybrid deep learning model for sentiment classification. IEEE Access 2020, 8, 58080–58093.

[11] A. Sharma and W. J. M. I. Verbeke, 2020. Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081). Front. Big Data, vol. 3, no. April, pp. 1–11. doi: 10.3389/fdata.2020.00015.

[12] G. Geetha, G. Saranya, K. Chakrapani, J. G. Ponsam, M. Safa, and S. Karpagaselvi, 2020. Early Detection of Depression from Social Media Data Using Machine Learning Algorithms. ICPECTS 2020 - IEEE 2nd Int. Conf. Power, Energy, Control Transm. Syst. Proc., pp. 3–8. doi: 10.1109 / ICPECTS49113 .2020. 9336974.

[13] S. Almouzini, M. Khemakhem, and A. Alageel., 2019. Detecting Arabic Depressed Users from Twitter Data. Procedia Comput. Sci., vol. 163, pp. 257–265. doi: 10.1016/j.procs.2019.12.107.

[14] O. B. Deho, W. A. Agangiba, F. L. Aryeh, and J. A. Ansah., 2018. Sentiment analysis with word embedding. IEEE Int. Conf. Adapt. Sci. Technol. ICAST, vol. 2018-Augus, pp. 1–4. doi: 10.1109/ICASTECH.2018.8506717

[15] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp., 2018. Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Comput. Sci., vol. 127, pp. 511–520. doi: 10.1016/j.procs.2018.01.150.

[16] Boser, B., Guyon, I., Vapnik, V., 1992. A Training Algorithm for Optimal Margin Classifiers. In: In ProceedIngs of the Fifth Annual Workshop on Computational LearnIng Theory, pp. 144–152.

[17] Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

[18] Sonal Singh, 2022. Leveraging Stacking Model to Identify Depression. INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 11, Issue 01 (January 2022),

[19] Z. Wang, H. Ren, R. Lu and L. Huang., 2022. Stacking Based LightGBM-CatBoost-RandomForest Algorithm and Its Application in Big Data Modeling. 4th International Conference on Data-driven Optimization of Complex Systems (DOCS), pp. 1-6, doi: 10.1109/DOCS55193.2022.9967714.

[20] Qiu, X.; Zhang, L.; Ren, Y.; Suganthan, P.N.; Amaratunga, G., 2014. Ensemble deep learning for regression and time series forecasting. In Proceedings of the 2014 IEEE symposium on Computational Intelligence in Ensemble Learning (CIEL), Orlando, FL, USA; pp. 1–6.

[21] Ankit and N. Saleena., 2018. An Ensemble Classification System for Twitter Sentiment Analysis. Procedia Comput. Sci., vol. 132, no. Iccids, pp. 937–946. doi: 10.1016/j.procs.2018.05.109

Authors' Profiles

Tejaswita Garg is research scholar at Jiwaji University, Gwalior, Madhya Pradesh. She has completed her Master of Technology in Computer Science from Banasthali Vidyapeeth, Rajasthan in 2013. She is interested in research areas including Machine Learning, Artificial Intelligence and Data Analytics.

Dr. Sanjay K. Gupta is Professor, Faculty of Computer Science and Applications, Jiwaji University, Gwalior, Madhya Pradesh. He has been associated in the academic activities of Jiwaji University and other Universities as well. He is a member of board of studies, faculty of Science, and Engineering of Jiwaji university, and of different institutions and academic activities. He is a life member of Professional Societies like Computer Society of India (CSI), YHAI, IAENG. Dr. Gupta has published more than 15 research papers in National and International Journals and Conferences. His research interest lies broadly in areas of Machine learning, Artificial Intelligence, IT applications, software engineering, and software testability of object-oriented systems.

How to cite this paper: Tejaswita Garg, Sanjay K. Gupta, "A Novel Algorithm for Stacked Generalization Approach to Predict Neurological Disorder over Digital Footprints", International Journal of Modern Education and Computer Science(IJMECS), Vol.15, No.5, pp. 60-73, 2023. DOI:10.5815/ijmecs.2023.05.05