Computers, Materials & Continua
DOI: 10.32604/cmc.2023.031867
Article

Sigmoidal Particle Swarm Optimization for Twitter Sentiment Analysis

Sandeep Kumar[1], Muhammad Badruddin Khan[2], Mozaherul Hoque Abul Hasanat[2], Abdul Khader Jilani Saudagar[2,*], Abdullah AlTameem[2] and Mohammed AlKhathami[2]

[1]Department of Computer Science and Engineering, CHRIST (Deemed to be University), Bangalore, 560074, India
[2]Information Systems Department, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, 11432, Saudi Arabia
*Corresponding Author: Abdul Khader Jilani Saudagar. Email: aksaudagar@imamu.edu.sa
Received: 28 April 2022; Accepted: 06 June 2022

Abstract: Social media, like Twitter, is a data repository, and people exchange views on global issues like the COVID-19 pandemic. Social media has been shown to influence the low acceptance of vaccines. This work aims to identify public sentiments concerning the COVID-19 vaccines and better understand the individual's sensitivities and feelings that lead to achievement. This work proposes a method to analyze the opinion of an individual's tweet about the COVID-19 vaccines. This paper introduces a sigmoidal particle swarm optimization (SPSO) algorithm. First, the performance of SPSO is measured on a set of 12 benchmark problems, and later it is deployed for selecting optimal text features and categorizing sentiment. The proposed method uses TextBlob and VADER for sentiment analysis, CountVectorizer, and term frequency-inverse document frequency (TF-IDF) vectorizer for feature extraction, followed by SPSO-based feature selection. The Covid-19 vaccination tweets dataset was created and used for training, validating, and testing. The proposed approach outperformed considered algorithms in terms of accuracy. Additionally, we augmented the newly created dataset to make it balanced to increase performance. A classical support vector machine (SVM) gives better accuracy for the augmented dataset without a feature selection algorithm. It shows that augmentation improves the overall accuracy of tweet analysis. After the augmentation performance of PSO and SPSO is improved by almost 7% and 5%, respectively, it is observed that simple SVM with 10-fold cross-validation significantly improved compared to the primary dataset.

Keywords: Twitter data analysis; sentiment analysis; social media analytics; swarm intelligence; COVID-19 vaccine

1 Introduction

The world experienced a difficult time during the last two years due to COVID-19. Most countries imposed complete lockdown, and people were forced to stay in their homes. This lockdown impacted the mental health of individuals around the globe. The good thing is that we now have a vaccine for

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

COVID-19. There was negative gossip about the vaccine for its side effects in the initial phase [1]. There were different rumors about COVID-19 vaccines, and all the governments faced a challenge in convincing people to vaccination. Twitter is a social media platform for sharing the opinion of an individual. Many Twitter users posted their views about COVID-19 vaccines at different stages. Thus, it is desirable to study the sentiments of people toward COVID-19 vaccines to prepare a plan for complete vaccination. Dores et al. [2] raised the issue that negative gossip may lead to social isolation.

Hayawi et al. [3] created a Twitter dataset to detect misinformation related to COVID-19 vaccination. Hayawi et al. [3] highlighted the requirement of sentiment analysis for English language tweets. This paper analyzed the view of individuals toward the COVID-19 vaccine and categorized them as positive, negative, and neutral. This paper highlights the various feature extraction, selection, and classification techniques used to analyze Twitter data and derive your opinions. Here, multiple features such as unigram and bigram are extracted to compare the precision of the methods. Additionally, some popular swarm intelligence-based approaches are deployed for selecting an optimal set of features.

Swarm intelligence (SI) algorithms successfully solved various complex optimization problems. These algorithms belong to the class of nature-inspired algorithms (NIA). The last three decades witnessed their exponentially increasing popularity due to their simplicity, flexibility, and broad applicability. Classification of NIAs is generally done based on the source of inspiration. These algorithms are based on biological phenomena (ex. genetic algorithm (GA) [4]), swarming behavior of birds (ex. particle swarm optimization (PSO) [5]), intelligent foraging (ex. ant colony optimization (ACO) [6], artificial bee colony (ABC) algorithm [7], bat algorithm (BA) [8] and many more), organisms-based (ex. cat swarm optimization [9]), social behavior-based [10–12].

This paper deployed a whale optimization algorithm (WOA) [13] and PSO for optimizing extracted features. With the help of exhaustive experiments, it is decided to improve PSO to get better performance. This paper presented a new variant of the PSO algorithm with an improved inertia weight strategy. The proposed approach was deployed for feature selection and improved training and testing accuracy.

The significant research contributions of this article are as follows:

- A new Twitter dataset was created for COVID-19 vaccine tweets. Newly created dataset uploaded on Github and publicly available for experiments. Link: https://github.com/sandpoonia/COVID-19-Vaccination-Tweets.
- A new variant of the PSO (Sigmoidal PSO) algorithm was developed and tested on benchmark problems.
- The sigmoidal PSO is deployed for feature selection with improved performance.
- Newly created dataset augmented and performance evaluated for all the considered algorithms.

The remaining paper is organized as follows: Section 2 discusses contemporary development in sentiment analysis and PSO algorithms. The new variant of PSO is explained in Section 3, and its performance is validated on benchmark problems in this section. Section 4 contains the proposed sentiment analysis model with detailed specifications of each step and analyzed results for the proposed model. Section 5 concludes the work done in this paper.

2 Related Literature

Melton et al. [14] analyze public sentiments for COVID-19 vaccines using the Reddit social media platform. Melton et al. [14] identified that peoples are more concerned about the side effects of vaccines. Sattar et al. [15] analyzed Twitter data for the USA and found that the public has healthy life after vaccination. Twitter is a social media platform from which data can be collected and used to analyze the sentiments of individuals. The computational approach to determining a tweet's nature (positive, negative, and neutral) is sentiment analysis (SA). Data gathered from Twitter is highly unstructured and ambiguous. SA first explores the text's sentiments and then extracts them, whereas opinion mining first extracts and analyzes them. SA is an approach for retrieving textual information that considers specific data analysis. It is a process of experiencing the sensations and viewpoints of people. Recently Tran et al. [16] conducted a student survey to analyze COVID-19 impact on their mental health.

2.1 Sentiment Analysis

Scientists and ordinary people posted many positive and negative stories after the announcement of the COVID-19 vaccines. These stories include the opinion of an individual about vaccine distribution, the side effects of vaccines, effectiveness, and many more. Hussain et al. [17] conducted an observational study to analyze the altitude of US and UK people towards COVID-19 vaccines by using Facebook and Twitter posts. They used lexicon-based analysis and a deep learning model for sentiment analysis. Hussain et al. [17] used a weighted average of VADERx and TextBlobx and combined it with the bidirectional encoder representations from transformers (BERT) model outcomes. Kwok et al. [18] deployed a machine learning technique to analyze Twitter sentiment relating to COVID-19 vaccination for Australia and found that people have mixed opinions.

Ritonga et al. [19] used a naive Bayes algorithm (NBA) to investigate the sentiments of Indonesian people regarding COVID-19 vaccination. This research found that more than half of the considered tweets are categorized as negative. Garcia et al. [20] analyzed news items related to COVID-19 for Brazil and organized them into ten categories based on ranking. Here, the author compared English and Portuguese tweets. Nurdeni et al. [21] considered two vaccines, Sinovac and Pfizer, to analyze individuals' sentiments from Indonesia. The model proposed by Nurdeni et al. [21] identified 77% and 81% positive sentiments for Sinovac and Pfizer, respectively. Manguri et al. [22] performed SA on the COVID-19 outbreak by collecting tweets worldwide. Gbashi et al. [23] analyzed news headlines and tweets regarding COVID-19 vaccination for the African continent. This study observed that very few users were active initially, which gradually increased with lockdown and vaccines' availability.

Sun et al. [24] developed a learning model for multiple features and highlighted the role of feature selection. Tran et al. [25] identified the role of domain-specific dictionaries and pre-defined rules in sentiment analysis. The author combined the attention method and rule-based approach for this purpose. Sun et al. [26] focused on feature optimization using a neural network. Bonnevie et al. [27] measured the opposition to the COVID-19 vaccine with the Twitter dataset. Recent research performed prediction [28], environmental [29], health [30,31], socioeconomic [32], emotional [33] impact of COVID-19. Iwendi et al. [34] suggested a new strategy to detect fake news related to COVID-19. The author deployed information fusion-based techniques for fetching news data and AI-based techniques for detection. This research identified that this is due to adverse health impacts, policies and politics, vaccine ingredients, clinical trials, and safety—most of the sentiment analysis is performed using machine learning (ML) and deep learning (DL) approaches. The performance of these techniques varies with the inclusion of optimization algorithms. This work considered the PSO algorithm for feature selection to improve the overall accuracy of classification. Tab. 1 discusses some more recent development in PSO.

Table 1: Recent modifications in PSO

Year	Author [Ref.]	Modification in PSO	Application	Remark
2017	Kiran [35]	Normal distribution based new position update	Engineering optimization	New position update strategy
2018	Tian et al. [36]	Chaos-based initialization and adaptive inertia weight	Image segmentation	A logistic map is deployed for uniform initialization
2018	Wang et al. [37]	Adaptive learning in PSO	Engineering optimization	A hybrid approach
2019	Ibrahim et al. [38]	Hybrid of salp swarm algorithm and PSO algorithms	Feature selection	A hybrid approach
2020	Zhang et al. [39]	Dynamic neighborhood-based learning approach used in PSO	Multimodal and multi-objective problems	Enhanced diversity
2020	Cui et al. [40]	Multi-objective version of PSO	Green coal production problem	Deployed for solving complex optimization problems
2020	Chen et al. [41]	PSO-based particle filter	Mechanical fault diagnosis	Introduced mutation operator in PSO
2020	El-Kenawy et al. [42]	Hybridized PSO with gray Wolf optimization algorithm	Feature selection	Tested over 17 datasets
2021	Wang et al. [43]	Mixed-variable encoding scheme	Mixed-variable optimization Problem	Continuous and discrete reproduction method
2021	Sedighizadeh et al. [44]	A dynamic inertia weight adjustment strategy proposed in PSO to improve exploration	Continuous space optimization	Proposed generalized PSO provides robust interrelation between particles

2.2 *Particle Swarm Optimization*

The PSO is a successful swarm intelligence-based algorithm in which a group of individuals explores the solicited resolution in the provided search space of the problem. The individuals are perceived as particles, and combinedly, they are perceived as a swarm. The PSO is an iterative non-deterministic stochastic algorithm. In PSO, every individual updates their position by learning from their previous best position (*pbest*) and global best position (*gbest*). Here the *gbest* is the position of the best-fit solution in the population. The fitness of the individuals is analyzed based on objective function value. Finally, the individuals conduct the solution search process by iteratively updating the given search space positions with the specific velocity [5].

Initially, the position and velocity of every individual are randomly initialized in the provided search space. In the next step, all the solutions revise their positions using the velocities while velocity is updated using Eq. (1).

$$V_{i+1,j} = W \times V_{ij} + \overbrace{ac_1 \times r_1 \times \left(pbest_{ij} - S_{ij}\right)}^{\text{Cognitive component}} + \overbrace{ac_2 \times r_2 \times \left(gbest_j - S_{ij}\right)}^{\text{Social component}} \tag{1}$$

In Eq. (1), current velocity is denoted by v_{ij}, the previous best solution of the current individual and best solution found so far are $pbest_{ij}$ and $gbest_j$, respectively, S_{ij} denotes the solution that will update its position, ac_1, ac_2 are the acceleration coefficients, and r_1, r_2 are the random number in the range (0, 1). Inertia weight, denoted by W, controls the velocity of an individual. Each solution updates its velocity by using Eq. (1).

$$\mathrm{S_{i+1,j}} = \mathrm{S_{ij}} + \overbrace{\mathrm{V_{i+1,j}}}^{\text{step}} \tag{2}$$

After that, the individuals update their positions using Eq. (2). In Eq. (2), $S_{i+1,j}$ is the updated position of the individual S_{ij}. In Eq. (2), i represents the individual who will update, whereas j shows the dimension D will be updated. The detailed pseudo-code of the PSO is described in Algorithm 1.

Algorithm 1: Particle Swarm Optimization

Initialize the position and velocity of N individuals.
Assign the values to the acceleration coefficients ac_1 and ac_2.
Evaluate the objective function value.
Identify the *gbest* and *pbest* from the population.
while Termination criteria are not met **do**
  for every individual, S_i **do**
    for each dimension j, S_{ij} **do**
      (i) Engender new velocity V_{ij} using (1).
      (ii) Engender new solution S_{ij} using (2).

    end for
  end for
  Evaluate the new solution.
  Update the *gbest* and *pbest* solutions.
end while
Return the best solution among the N individuals.

PSO has been modified numerous times for solving complex optimization problems as it is one of the most straightforward and robust swarm-based algorithms. Kiran [35] proposed a normal distribution-based new position update strategy in PSO that improved diversity in solutions. Eq. (3), μ denotes the mean, *and* σ represents the standard deviation.

$$x_{i,j}\left(t+1\right) = \mu + \sigma \times Z \tag{3}$$

Tian et al. [36] used a logistic map (refer to Eq. (4)) to change swarm initialization to generate uniform solutions. Modified inertia weight using the sigmoidal function to make it adaptive (refer to Eq. (5)) and improved diversity using wavelet mutation.

$$x_{n+1} = f(\mu, x_n) = \mu x_n (1 - x_n), n = 0, 1, 2, \ldots \tag{4}$$

$$\omega(t) = \left\{ \begin{array}{ll} 0.9, & t \leq \alpha t\_\{max\}\} \\ \dfrac{1}{1 + e^{10t - 2t_{max}/t_{max}}} + 0.4, & otherwise \end{array} \right\} \tag{5}$$

Recently Wang et al. [37,43] proposed a couple of new strategies in PSO. First, the adaptive learning approach [37] to avoid the problem of premature convergence, and the second approach was developed to solve an optimization problem with mixed variables. Zhang et al. [39] introduced a new variant with a dynamic neighborhood and successfully deployed it for multimodal and multi-objective problems.

Algorithm 2: Sigmoidal approach for inertia weight

Initialize iteration counter
Computer new inertia weight using following Equation
  $\omega_{iter} = 0.5 - 10^{log(iter)-2}$
if $\omega_{iter} \leq 0$, then
  $\omega_{iter} = 0.2$
else
  $\omega_{iter} = \omega_{iter-1}$
end if

3 Sigmoidal Particle Swarm Optimization

Exploitation and exploration are two main segments for a meta-heuristic algorithm to accomplish exact solutions and avoid trapping into local optima. Due to linearly decreasing inertia weight, sometimes PSO is trapped in the local solution as it shows poor diversity. Accordingly, the inertia weight is decided by the sigmoidal function in the anticipated variant. Due to the nonlinear nature of the sigmoidal function, it demonstrates improved results for optimization. In SPSO, the inertia weight is changed with sigmoidal function. Inertia weight plays a vital role in controlling the convergence in PSO. Initially, it was kept constant, but with experiments, it is observed that decreasing inertia weight improves the overall solution quality. The new approach decides it with sigmoidal nonlinear function within the range [0.5, 0.2]. The proposed approach computes inertia weight using Algorithm 2. In Algorithm 2, ω_{iter} is inertia weight for iteration *iter*. The upper and lower bound for parameter ω_{iter} is decided empirically.

The performance of SPSO is tested over a set of 12 standard benchmark problems and compared with basic PSO, differential evolution (DE) [45], and ABC algorithms in terms of success rate (SR) and the average number of function evaluations (AFE). The considered problems are given in Tab. 2. The designated problems are different and with varying complexity levels. The experiments are performed on MATLAB R2020b using Intel core i7 machine with 16 GB RAM and 8GB NVIDIA GTX graphics processor. Results are shown in Tab. 3 for considered algorithms. Parameter settings for all the algorithms are taken from their base papers. Results prove that SPSO outperforms all the considered algorithms for ten problems in terms of AFE and SR.

Table 2: Benchmark problems

Equation	Range	Optimal value
$F_1(X) = \sum_{i=1}^{d} x_i^2$	[−100, 100]	0
$F_2(X) = \sum_{i=1}^{n} i \times (x_i)^4$	[−5.12, 5.12]	0
$F_3(X) = \frac{1}{4000}\left(\sum_{i=1}^{D}(x_i^2)\right) - \left(\prod_{i=1}^{D} cos\left(\frac{x_i}{\sqrt{i}}\right)\right) + 1$	[−600, 600]	0
$F_4(X) = \sum_{i=1}^{D}\left[x_i^2 - 10cos(2\pi x_i) + 10\right]$	[−5.12, 5.12]	0
$F_5(X) = exp\left(-0.5\sum_{i=1}^{n} x_i^2\right)$	[−1, 1]	1
$F_6(X) = x_1^2 + 10^6 \sum_{i=2}^{n} x_i^2$	[−10, 10]	0
$F_7(X) = 1 - cos(2\pi p) + 0.1 \times p, where, p = \sqrt{\sum_{i=1}^{D} x_i^2}$	[−100, 100]	0
$F_8(X) = \sum_{i=1}^{D} ix_i^2$	[−5.12, 5.12]	0
$F_9(X) = \sum_{i=1}^{D} \lvert x_i\rvert^{i+1}$	[−1, 1]	0
$F_{10}(X) = \sum_{i=1}^{D} (\lvert x_i + 0.5\rvert)^2$	[−100, 100]	0
$F_{11}(X) = \sum_{i=1}^{d} ix_i^4 + random[0, 1)$	[−1.28, 1.28]	0
$F_{12}(X) = \sum_{i=1}^{D}\sum_{j=1}^{i} x_j^2$	[−65.536, 65.536]	0

Table 3: Comparison of AFE and SR for PSO and SPSO

Function	Dim	PSO		DE		ABC		SPSO	
		AFE	SR	AFE	SR	AFE	SR	AFE	SR
F_1	30	41404	100	23226	100	23218	100	8793	100
F_2	30	50391	100	21337.5	100	10977.5	100	10288.5	100
F_3	30	45565.5	100	64616	81	76412	68	9875	100
F_4	30	200050	0	200050	0	99490	2	198865	6
F_5	30	54482	100	17692	100	100007.87	100	14784	100
F_6	30	56186.5	97	40413.5	100	41584	100	24938.5	97
F_7	4	2181.5	100	21289	100	99967.23	1	986	100
F_8	2	1419.5	100	26440	100	23131	100	612.5	100

(Continued)

Table 3: Continued

Function	Dim	PSO		DE		ABC		SPSO	
		AFE	SR	AFE	SR	AFE	SR	AFE	SR
F_9	2	1842.5	100	196707	2	100000.08	0	827	100
F_{10}	3	1125.5	100	8735	100	22335	100	603.5	100
F_{11}	6	73443.5	65	33200	91	18030.84	100	33752	86
F_{12}	4	5267.5	100	200050	0	100035.22	0	1807	100

The analysis of convergence speed was performed using acceleration rate (AR). AR is computed using Eq. (6). SPSO has high AR if it takes fewer iterations to find the optimal solution. Thus, low AFE for the proposed algorithm leads to a high acceleration rate or fast convergence.

$$AR = \frac{AFE_{Algo}}{AFE_{SPSO}}, \ Where\ Algo \in \{DE, ABC, PSO\} \tag{6}$$

Tab. 4 illustrates the comparison of SPSO with PSO, DE, and ABC for AR. SPSO is better than these algorithms in terms of AR except f_4 and f_{11}.

Table 4: Comparison of AR for SPSO

Algorithm\Function	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
PSO	4.7	4.9	4.6	1	3.7	2.3	2.21	2.32	2.23	1.86	2.2	2.92
DE	2.6	2.1	6.5	1	1.2	1.6	21.59	43.2	237.9	14.5	1	110.7
ABC	2.6	1.1	7.7	0.5	6.8	1.7	101.4	37.8	120.9	37	0.5	55.36

Furthermore, results are compared using boxplot graphs. Boxplot graph tells us about the distribution of data. Fig. 1 shows that SPSO takes less AFE to get an optimal result. Even the median for SPSO is less than the first quartile of DE and ABC.

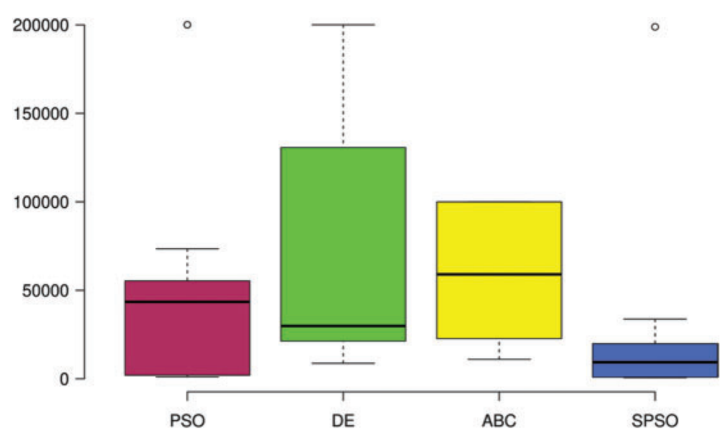

Figure 1: Boxplots graph for AFE

4 Twitter Sentiment Analysis Using SPSO-based Bag-of-Words

This paper introduced optimal bag-of-words with sigmoidal PSO for tweet sentiment analysis and classified them into three standard categories, negative, positive, and neutral. This tweet classification method works on five steps, as depicted in Fig. 2. The first step retrieves tweets related to Covid-19 vaccination using the Tweepy application programming interface (API). The second stage uses text pre-processing techniques such as stop words, tokenization, and lemmatization and joins in cleaning up the text data. Phase III uses Counter Vectorizer and TF-IDF Vectorizer to extract features from Covid-19 vaccination tweets. The fourth step uses modified PSO for selecting tweet features based on training fitness. Finally, classify the Covid-19 vaccination Twitter post as positive, negative, and neutral. This paper uses the Covid-19 vaccination tweet dataset for training and validating the proposed model.

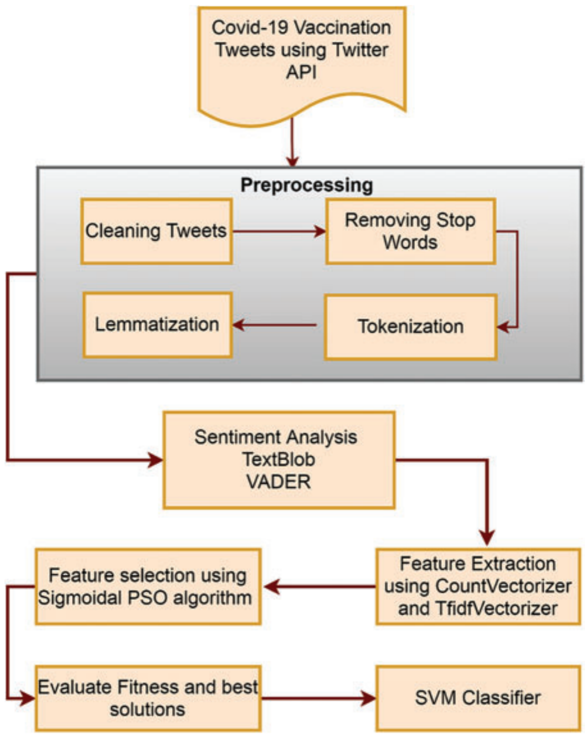

Figure 2: Workflow for sentiment analysis

4.1 Data Gathering

Twitter posts are collected from the social media platform Twitter developer account using Tweepy API [46,47] method to get Covid-19 vaccination public opinion. We created a Twitter developer account linked to a Twitter account and used consumer keys and authentication tokens to retrieve tweets. The retrieved tweets have different fields, namely tweet_text, tweet_created_at, tweet_source, tweet_location, tweet_like, and tweet_retweet. These tweets mainly belong to USA, India, UK, Canada, Australia, and Russia. This work mainly focused on sentiment analysis for four COVID-19 vaccines: Covaxin, Covishield, Sputnik, and Pfizer. We filtered out tweets that mentioned Covid-19 vaccination, Covishield, Covaxin, Pfizer, and Sputnik in these hashtags.

Tab. 5 illustrated the detail of retweets and likes for collected tweets and reported mean, median, and IQR for these two features. Here, the mean indicates the average number of retweets/likes, which indicates that the total number of tweets is low, but they have a significant impact and reflect public opinion because of the large number of retweets. The mean value (mean) for retweets/likes is also high enough. The interquartile range (IQR) shows the spread of the data.

Table 5: Descriptive statistics of collected Twitter data

Number of tweets	PSO			DE		
	Mean	Median	IQR	Mean	Median	IQR
9799	40883	308	1814	12693	1311	8179

4.2 Data Preprocessing

The first step of the proposed model is to preprocess Twitter post that brings tweets into a specific form that is analyzable and predictable. Textual data preprocessing, needs to apply some steps to transform text data into numerical features, and these steps are dependent on the domain of the data itself. For the Covid-19 vaccination tweet dataset, apply the five steps for preprocessing as discussed below.

Tweet Cleaning: The first step in text cleaning is to remove unwanted characters from each tweet. In this step, eliminate text and characters which are irrelevant noise from these tweets. For example, eliminate *URLs, # from hashtags, HTTPS, parenthesis, slashes, and @Username* using regular expression (RE) [48]. Replace multiple spaces with a single space and eliminate special symbols and characters. Once text cleaning is completed, tweets are ready for the next step to analyze the sentiment of each tweet.

Remove Stop words: English words do not add much meaning to any Twitter post and are filtered out before processing these tweets [49]. Remove these stop words from cleaned tweets for further processing. Tokenization and Lemmatization: Tokenization splits each tweet into a smaller unit as an individual word which is a token. This is a process that protects sensitive data using algorithmically generated token numbers. Different forms of a similar token word are grouped using lemmatization. It removes the modulate ending of each token and returns it to a word's base form.

4.3 Sentiment Analysis

This is an opinion mining process that determines the attitude or emotion of each tweet as positive, negative, or neutral [50]. Sentiment analysis is performed on cleaned tweets to understand and find the opinion on the Covid-19 vaccination. This work used two methods to analyze tweet attitudes: TextBlob [51] and VADER. TextBlob function finds the subjectivity and polarity of each tweet. Subjectivity refers to the opinion of a tweet, which lies in the range of [0, 1] and refers to personal opinion or emotion. Polarity finds the tweet sentiment analysis in [-1, 1]. Here,-1 represents negative, and 1 represents positive sentiment. VADER is a rule-based analysis tool that rates each tweet positively, negatively, and neutral and finds the tweet's overall compound rating. Based on sentiment analysis and human annotation, it was found that this dataset had 56% positive, 37% negative, and 8% neutral tweets.

4.4 Feature Extraction

Twitter post data are not computable directly, so they should be transformed into numerical data as vector space using the feature extraction method. Bag-of-words is a representative data model used for feature extraction and count frequency of each word in a document [52,53]. This is usually used for clustering and classification. The proposed method uses two methods for feature extraction: Counter Vectorizer and TF-IDF. Counter Vectorizer is used to transform each tweet into a vector based on the frequency of each word for the entire tweet dataset. TF-IDF is a popular technique for information retrieval and analyzes essential words in a document. This method considers the critical words and skips commonly used words. Extract features with a different range of n-grams and passes to the next feature selection step.

4.5 Feature Selection

The features extracted from the earlier step are passed to the SPSO algorithm to select the valuable features. This phase selects the notable suitable features that help predict the class of each tweet. Subsequently, the features extracted from the last step are used to determine the optimal features and form clusters using the particular features, improving accuracy and reducing over-fitting. Here sigmoidal PSO is used for clustering to choose the optimal set of features.

4.6 Classification

Classification of Covid-19 vaccination tweets is the final step of this model. First, the selected features from the earlier step are passed to SVM Classifier. Then, SVM is trained based on the input features vector and predicts the output. Once the labeling of the training dataset is done, a test dataset tests the accuracy of the proposed model. The complete process has six functions: tweet fetching, data pre-processing, sentiment analysis, feature extraction, feature selection, and classifying of training and testing datasets to find accuracy.

4.7 Result Analysis of Feature Selection Technique for the Original Dataset

The first step determined the sentiments from TextBlob and VADER's sentiment analyzers. Once tweet sentiment analysis is completed, it determines the accuracy of the proposed model; this section tests the performance of modified PSO and compares it with other algorithms for the COVID-19 vaccination tweets dataset collected from Twitter through Twitter tweepy API. The proposed method is used for feature selection to improve fitness. This dataset has 9799 tweets for training and testing the proposed method. Here, a randomly selected 25% dataset (i.e., 2450 tweets) is used for testing, and the rest, 75%, is used for training the proposed model. The proposed tweet analysis method has been compared with WOA and PSO algorithms. The confusion matrix describes the performance of each model. The confusion matrix for the count vectorizer is depicted in Fig. 3, and for TF-IDF is depicted in Fig. 4. All the algorithms use an equal number of tweets. The performance is measured for fitness, selected features, training, and test accuracy. It measures precision, F1 score, recall, macro average, and the weighted average for the positive, negative and neutral dataset. Also, dealing with training accuracy and test accuracy, the comparison is depicted in Tab. 6. It is observed that the proposed method shows a better result than other existing algorithms.

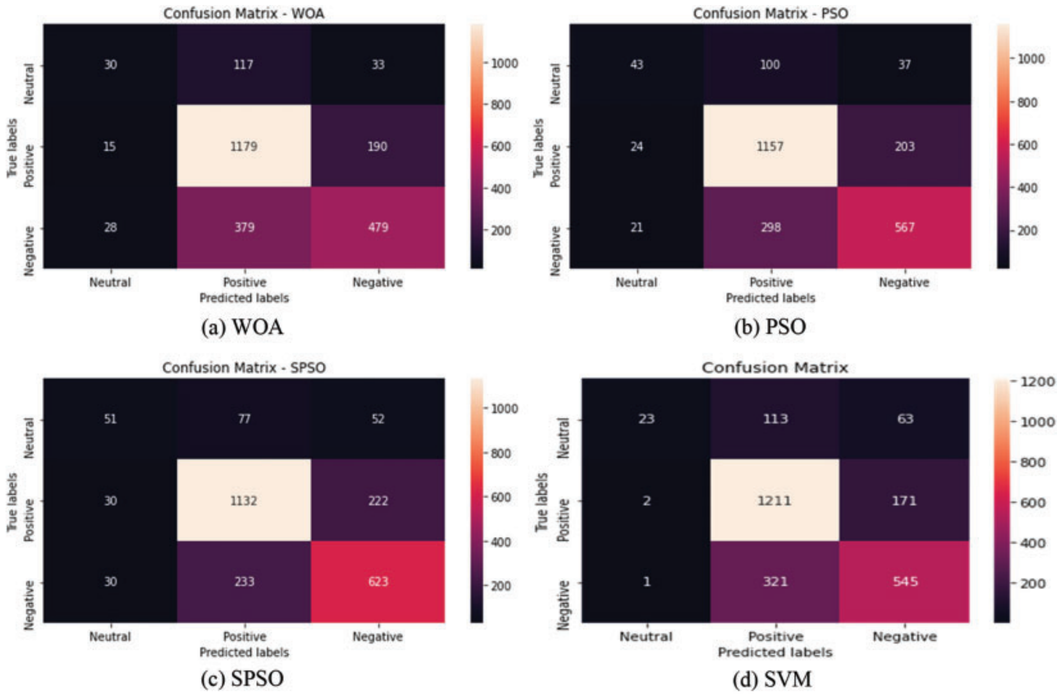

Figure 3: Confusion matrix for count vectorizer method

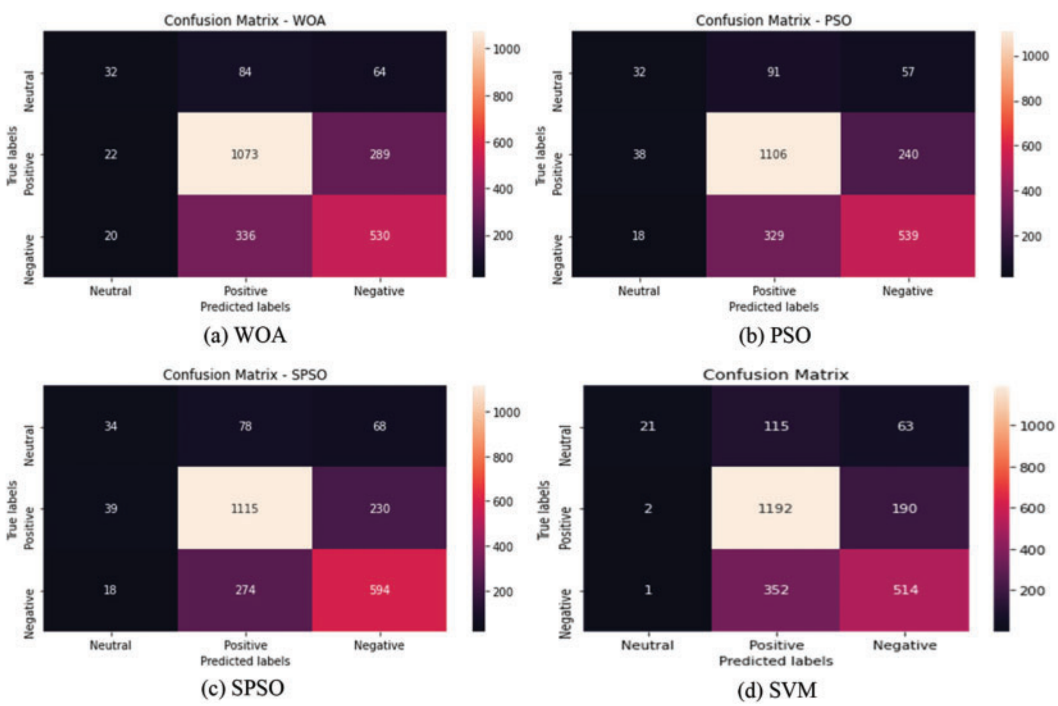

Figure 4: Confusion matrix for TF-IDF vectorizer method

Table 6: Performance analysis of SPSO for the original dataset

Algorithm	Parameter	Count vectorizer			TF-IDF vectorizer			Support
		Precision	Recall	F1–score	Precision	Recall	F1–score	
WOA + SVM	Neutral	0.41	0.17	0.24	0.43	0.18	0.25	180
	Positive	0.7	0.85	0.77	0.72	0.78	0.75	1384
	Negative	0.68	0.54	0.6	0.6	0.6	0.6	886
	Macro avg	0.6	0.52	0.54	0.58	0.52	0.53	2450
	Weighted avg	0.67	0.69	0.67	0.65	0.67	0.66	2450
	Accuracy	-	-	0.69	-	-	0.67	2450
SVM (With feature selection algorithm)	Neutral	0.88	0.12	0.2	0.88	0.11	0.19	180
	Positive	0.74	0.88	0.8	0.72	0.86	0.78	1384
	Negative	0.7	0.63	0.66	0.67	0.59	0.63	886
	Macro avg	0.77	0.54	0.56	0.75	0.52	0.53	2450
	Weighted avg	0.74	0.73	0.7	0.71	0.7	0.68	2450
	Accuracy	-	-	0.73	-	-	0.7	2450
PSO + SVM	Neutral	0.49	0.24	0.32	0.36	0.18	0.24	180
	Positive	0.74	0.84	0.79	0.72	0.8	0.76	1384
	Negative	0.7	0.64	0.67	0.64	0.61	0.63	886
	Macro avg	0.65	0.57	0.59	0.58	0.53	0.54	2450
	Weighted avg	0.71	0.72	0.71	0.67	0.68	0.67	2450
	Accuracy	-	-	0.72	-	-	0.68	2450
SPSO + SVM	Neutral	0.46	0.28	0.35	0.37	0.19	0.25	180
	Positive	0.79	0.82	0.8	0.76	0.81	0.78	1384
	Negative	0.69	0.7	0.7	0.67	0.67	0.67	886
	Macro avg	0.65	0.6	0.62	0.6	0.55	0.57	2450
	Weighted avg	0.73	0.74	0.73	0.7	0.71	0.7	2450
	Accuracy	-	-	0.74	-	-	0.71	2450

4.8 Result Analysis for Augmented Dataset with 10-fold Cross-validation

The present dataset is imbalanced as there is a huge difference in neutral and positive tweet count. An imbalanced dataset leads to less accurate prediction and classification. In order to balance the created dataset, the author performed augmentation for neutral tweets. As a result, the new dataset contains 26% neutral tweets, while positive and negative tweets are 45% and 29%, respectively. Therefore, the augmented dataset is balanced in terms of tweets in all three categories. The summary of the augmented dataset is depicted in Tab. 7.

Table 7: Performance analysis of SPSO for augmented dataset

Algorithm	Parameter	Count vectorizer			TF-IDF vectorizer			Support
		Precision	Recall	F1–score	Precision	Recall	F1–score	
SPSO + SVM	Neutral	0.82	0.9	0.86	0.76	0.84	0.8	3153
	Positive	0.73	0.85	0.79	0.71	0.82	0.76	5497

(Continued)

Table 7: Continued

Algorithm	Parameter	Count vectorizer			TF-IDF vectorizer			Support
		Precision	Recall	F1–score	Precision	Recall	F1–score	
	Negative	0.78	0.52	0.63	0.75	0.49	0.59	3609
	Macro avg	0.78	0.76	0.76	0.74	0.72	0.72	12259
	Weighted avg	0.77	0.77	0.76	0.73	0.73	0.72	12259
	Accuracy	-	-	0.77	-	-	0.73	12259
PSO + SVM	Neutral	0.93	0.9	0.91	0.79	0.86	0.82	3153
	Positive	0.72	0.89	0.79	0.73	0.83	0.78	5497
	Negative	0.78	0.51	0.62	0.78	0.54	0.64	3609
	Macro avg	0.81	0.77	0.77	0.76	0.74	0.75	12259
	Weighted avg	0.79	0.78	0.77	0.76	0.75	0.75	12259
	Accuracy	-	-	0.78	-	-	0.75	12259
SVM (With feature selection algorithm)	Neutral	0.93	0.96	0.95	0.87	0.96	0.91	3153
	Positive	0.79	0.88	0.83	0.79	0.87	0.83	5497
	Negative	0.81	0.63	0.71	0.82	0.62	0.71	3609
	Macro avg	0.84	0.83	0.83	0.83	0.82	0.82	12259
	Weighted avg	0.83	0.83	0.83	0.82	0.82	0.81	12259
	Accuracy	-	-	0.83	-	-	0.82	12259

This subsection highlights the endorsement of sentiment analyzers, optimization algorithms, and machine learning algorithms to determine training and testing accuracy rates for the augmented dataset of COVID-19 vaccination sentiments. This augmented dataset has 12259 tweets for training and testing the proposed method. The augmented dataset was analyzed by SVM with 10-fold cross-validation.

10-fold cross-validation applied and augmented dataset partitioned into ten equal-sized sub-samples. Finally, the result of all ten folds is combined to form a single result. The best part of this approach is that it gives a chance to each sample to participate in the training and testing phase. The process of cross-validation iterated 10-times with each sub-sample used as validation. The confusion matrix describes the performance of each model. All the algorithms use an equal number of tweets. The performance is measured for fitness, selected features, training, and test accuracy. It measures precision, F1 score, recall, macro average, and the weighted average for the positive, negative and neutral dataset. Here, 10-fold cross-validation randomly takes the 10th part of the dataset (i.e., 1226 tweets out of 12259 tweets) used for testing, and the rest are used for training during each fold. Also, dealing with accuracy, the comparison is depicted in Tab. 7. It is observed that the proposed method shows a better result for imbalanced datasets, and simple SVM outperformed other algorithms for the augmented dataset. Results demonstrate that SVM classifies most of the time correctly, and there is more than a 13% increase in accuracy after augmentation. Thus, data augmentation improves performance for imbalanced datasets, and it does not require an exclusive feature selection mechanism.

Reviews about publicly available COVID-19 vaccines on Twitter were initially driven by the best COVID-19 vaccines and revealed active news topics in the mainstream media. The government can frame policies for sharing the ingredients of vaccines and discussing their side effects.

5 Conclusion

This paper proposed a new variant of PSO with a new mechanism for inertia weight calculation. Inertia weight computed with nonlinear sigmoidal function. The proposed approach is named sigmoidal PSO. The sigmoidal PSO was tested over twelve benchmark problems and deployed for Twitter sentiment analysis, and it gives better results than basic PSO and other considered algorithms. The sentiment analysis was carried out over a user-created Twitter dataset for COVID-19 vaccinations. It is observed that almost 56% of people have a positive attitude toward newly developed vaccines. Results show that SPSO gives 74% and 71% accuracy for count vectorizer and TF-IDF, respectively, significantly higher than the considered algorithms.

Additionally, augmentation was performed to balance the dataset, and 10-fold cross-validation was employed for the sentiment analysis over the augmented dataset. Results proved that sigmoidal PSO performed better for imbalanced datasets, while classical SVM gives better results for augmented datasets with a 5% improvement in accuracy compared to SPSO. In the future, sigmoidal PSO may be deployed for the image dataset. Furthermore, the dataset prepared for COVID-19 vaccination may be extended to analyze the mental health of individuals during and after the COVID-19.

Funding Statement: This research was supported by Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia, for funding this research work through project number 959.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. S. Islam, A. H. M. Kamal, A. Kabir, D. L. Southern, S. H. Khan *et al.,* “Covid-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence,” *PloS One*, vol. 16, no. 5, 2021. https://doi.org/10.1371/journal.pone.0251605.

[2] T. D. Dores Cruz, R. van der Lee and B. Beersma, “Gossip about coronavirus: Infection status and norm adherence shape social responses,” *Group Processes & Intergroup Relations*, vol. 24, no. 4, pp. 658–679, 2021.

[3] K. Hayawi, S. Shahriar, M. A. Serhani, I. Taleb and S. S. Mathew, “Anti-vax: A novel twitter dataset for covid-19 vaccine misinformation detection,” *Public Health*, vol. 203, pp. 23–30, 2021.

[4] J. H. Holland and J. S. Reitman, “Cognitive systems based on adaptive algorithms,” in *Pattern Directed Inference Systems*, Orlando, Florida: Academic Press, Elsevier, pp. 313–329, 1978.

[5] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proc*. in *of ICNN’95-Int. Conf. on Neural Networks*, Perth, WA, Australia, IEEE, vol. 4, pp. 1942–1948, 1995.

[6] M. Dorigo and G. D. Caro, “Ant colony optimization: A new meta-heuristic,” in *Proc*. in *of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406)*, Washington, DC, USA, IEEE, vol. 2, pp. 1470–1477, 1999.

[7] D. Karaboga, “An idea based on honey bee swarm for numerical optimization,” *Technical Report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department*, vol. 200, pp. 1–10, 2005.

[8] X. S. Yang, “Bat algorithm for multi-objective optimization,” *International Journal of Bio-Inspired Computation*, vol. 3, no. 5, pp. 267–274, 2011.

[9] S. C. Chu, P. W. Tsai and J. S. Pan, “Cat swarm optimization,” in *Proc. of 9th Pacific Rim Int. Conf. on Artificial Intelligence*, Guilin, China, Springer, pp. 854–858, 2006.

[10] J. C. Bansal, H. Sharma, S. S. Jadon and M. Clerc, “Spider monkey optimization algorithm for numerical optimization,” *Memetic Computing*, vol. 6, no. 1, pp. 31–47, 2014.

[11] S. Mirjalili, S. M. Mirjalili and A. Hatamlou, “Multi-verse optimizer: A nature-inspired algorithm for global optimization,” *Neural Computing and Applications*, vol. 27, no. 2, pp. 495–513, 2016.
[12] S. Mirjalili, S. M. Mirjalili and A. Lewis, “Grey wolf optimizer,” *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.
[13] S. Mirjalili and A. Lewis, “The whale optimization algorithm,” *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016.
[14] C. A. Melton, O. A. Olusanya, N. Ammar and A. Shaban-Nejad, “Public sentiment analysis and topic modeling regarding covid-19 vaccines on the reddit social media platform: A call to action for strengthening vaccine confidence,” *Journal of Infection and Public Health*, vol. 14, no. 10, pp. 1505–1512, 2021.
[15] N. S. Sattar and S. Arifuzzaman, “Covid-19 vaccination awareness and aftermath: Public sentiment analysis on twitter data and vaccinated population prediction in the USA,” *Applied Sciences*, vol. 11, no. 13, pp. 6128, 2021.
[16] T. K. Tran, H. Dinh, H. Nguyen, D. N. Le, D. K. Nguyen *et al.,* “The impact of the COVID-19 pandemic on college students: An online survey,” *Sustainability*, vol. 13, no. 19, pp. 1–19, 2021.
[17] A. Hussain, A. Tahir, Z. Hussain, Z. Sheikh, M. Gogate *et al.,* “Artificial intelligence-enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the United Kingdom and the United States: Observational study,” *Journal of Medical Internet Research*, vol. 23, no. 4, 2021. https://doi.org/10.2196/26627.
[18] S. W. H. Kwok, S. K. Vadde and G. Wang, “Tweet topics and sentiments relating to covid-19 vaccination among Australian twitter users: Machine learning analysis,” *Journal of Medical Internet Research*, vol. 23, no. 5, 2021. https://doi.org/10.2196/26953.
[19] M. Ritonga, M. A. Al Ihsan, A. Anjar and F. H. Rambe, “Sentiment analysis of covid-19 vaccine in Indonesia using naïve Bayes algorithm,” in *Proc. Annual Conf. on Computer Science and Engineering Technology*, Medan, Indonesia, vol. 1088, no. 1, 2021.
[20] K. Garcia and L. Berton, “Topic detection and sentiment analysis in twitter content related to covid-19 from Brazil and the USA,” *Applied Soft Computing*, vol. 101, 2021. https://doi.org/10.1016/j.asoc.2020.107057.
[21] D. A. Nurdeni, I. Budi and A. B. Santoso, “Sentiment analysis on covid19 vaccines in Indonesia: From the perspective of sinovac and pfizer,” in *Proc. 2021 3rd East Indonesia Conf. on Computer and Information Technology (EIConCIT)*, Surabaya, Indonesia, IEEE, pp. 122–127, 2021.
[22] K. H. Manguri, R. N. Ramadhan and P. R. M. Amin, “Twitter sentiment analysis on worldwide covid-19 outbreaks,” *Kurdistan Journal of Applied Research*, vol. 5, no. 3, pp. 54–65, 2020.
[23] S. Gbashi, O. A. Adebo, W. Doorsamy and P. B. Njobeh, “Systematic delineation of media polarity on covid-19 vaccines in Africa: Computational linguistic modeling study,” *JMIR Medical Informatics*, vol. 9, no. 3, 2021. https://doi.org/10.2196/22916.
[24] W. Sun, X. Chen, X. Zhang, G. Dai, P. Chang *et al.,* “A multi-feature learning model with enhanced local attention for vehicle re-identification,” *Computers, Materials & Continua*, vol. 69, no. 3, pp. 3549–3561, 2021.
[25] T. K. Tran and T. T. Phan, “Capturing contextual factors in sentiment classification: An ensemble approach,” *IEEE Access*, vol. 22, no. 8, pp. 116856–116865, 2020.
[26] W. Sun, G. Zhang, X. Zhang, X. Zhang and N. Ge, “Fine-grained vehicle type classification using lightweight convolutional neural network with feature optimization and joint learning strategy,” *Multimedia Tools and Applications*, vol. 80, no. 20, pp. 30803–30816, 2021.
[27] E. Bonnevie, A. Gallegos-Jeffrey, J. Goldbarg, B. Byrd and J. Smyser, “Quantifying the rise of vaccine opposition on twitter during the covid-19 pandemic,” *Journal of Communication in Healthcare*, vol. 14, no. 1, pp. 12–19, 2020.
[28] N. Jain, S. Jhunthra, H. Garg, V. Gupta, S. Mohan *et al.,* “Prediction modelling of COVID using machine learning methods from B-cell dataset,” *Results in Physics*, vol. 21, 2021. https://doi.org/10.1016/j.rinp.2021.103813.
[29] V. Gupta, N. Jain, D. Virmani, S. Mohan, A. Ahmadian *et al.,* “Air and water health: Industrial footprints of COVID-19 imposed lockdown,” *Arabian Journal of Geosciences*, vol. 15, no. 8, pp. 1–8, 2022.

[30] S. Mohan, A. Abugabah, S. S. Kumar, A. K. Bashir and L. Sanzogni, "An approach to forecast impact of covid-19 using supervised machine learning model," *Software: Practice and Experience*, vol. 52, no. 4, pp. 824–840, 2022.

[31] C. Iwendi, C. G. Huescas, C. Chakraborty and S. Mohan, "COVID-19 health analysis and prediction using machine learning algorithms for Mexico and Brazil patients," *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–21, 2022. https://doi.org/10.1080/0952813X.2022.2058097.

[32] V. Gupta, K. C. Santosh, R. Arora, T. Ciano, K. S. Kalid *et al.,* "Socioeconomic impact due to COVID-19: An empirical assessment," *Information Processing & Management*, vol. 59, no. 2, 2022. https://doi.org/10.1016/j.ipm.2021.102810.

[33] V. Gupta, N. Jain, P. Katariya, A. Kumar, S. Mohan *et al.,* "An emotion care model using multimodal textual analysis on COVID-19," *Chaos, Solitons & Fractals*, vol. 144, 2021. https://doi.org/10.1016/j.chaos.2021.110708.

[34] C. Iwendi, S. Mohan, E. Ibeke, A. Ahmadian and T. Ciano, "Covid-19 fake news sentiment analysis," *Computers and Electrical Engineering*, vol. 101, 2022. https://doi.org/10.1016/j.compeleceng.2022.107967.

[35] M. S. Kiran, "Particle swarm optimization with a new update mechanism," *Applied Soft Computing*, vol. 60, pp. 670–678, 2017.

[36] D. Tian and Z. Shi, "MPSO: Modified particle swarm optimization and its applications," *Swarm and Evolutionary Computation*, vol. 41, pp. 49–68, 2018.

[37] F. Wang, H. Zhang, K. Li, Z. Lin, J. Yang *et al.,* "A hybrid particle swarm optimization algorithm using adaptive learning strategy," *Information Sciences*, vol. 436, pp. 162–177, 2018.

[38] R. A. Ibrahim, A. A. Ewees, D. Oliva, M. A. Elaziz and S. Lu, "Improved salp swarm algorithm based on particle swarm optimization for feature selection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 8, pp. 3155–3169, 2019.

[39] X. W. Zhang, H. Liu and L. P. Tu, "A modified particle swarm optimization for multimodal multi-objective optimization," *Engineering Applications of Artificial Intelligence*, vol. 95, pp. 103905, 2020.

[40] Z. Cui, J. Zhang, D. Wu, X. Cai, H. Wang *et al.,* "Hybrid many-objective particle swarm optimization algorithm for green coal production problem," *Information Sciences*, vol. 518, pp. 256–271, 2020.

[41] H. Chen, D. L. Fan, L. Fang, W. Huang, J. Huang *et al.,* "Particle swarm optimization algorithm with mutation operator for particle filter noise reduction in mechanical fault diagnosis," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 34, no. 10, 2020. https://doi.org/10.1142/S0218001420580124.

[42] E. S. El-Kenawy and M. Eid, "Hybrid gray wolf and particle swarm optimization for feature selection," *International Journal of Innovative Computing Information and Control*, vol. 16, no. 3, pp. 831–844, 2020.

[43] F. Wang, H. Zhang and A. Zhou, "A particle swarm optimization algorithm for mixed variable optimization problems," *Swarm and Evolutionary Computation*, vol. 60, 2021. https://doi.org/10.1016/j.swevo.2020.100808.

[44] D. Sedighizadeh, E. Masehian, M. Sedighizadeh and H. Akbaripour, "GEPSO: A new generalized particle swarm optimization algorithm," *Mathematics and Computers in Simulation*, vol. 179, pp. 194–212, 2021.

[45] R. Storn and K. Price, "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[46] Tweepy, "Tweepy," 2018. [Online]. Available: https://www.tweepy.org/.

[47] J. Roesslein, "Tweepy documentation," 2009. [Online]. Available: http://docs.tweepy.org/en/v3.5.0/.

[48] S. Gai and D. Malagrino, "System and method for performing regular expression matching with high parallelism," *US Patent*, vol. 7, no. 225, pp. 188, 2007.

[49] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of Information Science*, vol. 18, no. 1, pp. 45–55, 1992.

[50] E. C. Dragut, C. Yu, P. Sistla and W. Meng, "Construction of a sentimental word dictionary," in *Proc. of the 19th ACM Int. Conf. on Information and Knowledge Management*, Toronto ON Canada, pp. 1761–1764, 2010.

[51] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh *et al.,* “Textblob: Simplified text processing,” *Secondary TextBlob: Simplified Text Processing*, vol. 3, 2014.
[52] K. S. Kalaivani, S. Uma and C. S. Kanimozhiselvi, “A review on feature extraction techniques for sentiment classification,” in *Proc. of 2020 Fourth Int. Conf. on Computing Methodologies and Communication (ICCMC)*, Erode, India, IEEE, pp. 679–683, 2020.
[53] M. Z. Asghar, A. Khan, S. Ahmad and F. M. Kundi, “A review of feature extraction in sentiment analysis,” *Journal of Basic and Applied Scientific Research*, vol. 4, no. 3, pp. 181–186, 2014.