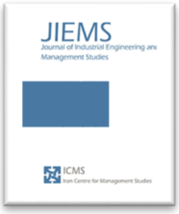
Vol.10, No.1, 2023, Pages 1-15

JIEMS

Journal of Industrial Engineering and Management Studies

journal homepage: www.jiems.icms.ac.ir


JIEMS
Journal of Industrial Engineering and Management Studies
ICMS
Iran Centre for Management Studies

Hybrid PSO-GSA based approach for feature selection

Monireh Hosseini [*1], Mahjoob Sadat Navabi [1]

[1] Department of Information Technology, Faculty of Industrial Engineering, K. N. Toosi University of Technology, Tehran, Iran.

Received: Aug 2022-16/ Revised: Dec 2022-29/ Accepted: Jan 2023-16

Abstract

With the development and widespread use of social networks among people, high-volume data is produced and the analysis of this data can be useful in many areas, including people's daily lives. Classification of this volume of data using traditional methods is a very difficult, time-consuming, and low-accuracy task, therefore, using sentiment analysis techniques, people's opinions can be effectively summarized and categorized. To this end, we propose an algorithm that combines Particle Swarm Optimization (PSO) and Gravitational Search Algorithm (GSA). The reason for combining the two algorithms is that the GSA has a good ability to search overall, but in the last iterations, it has a low speed in exploiting the search space. Since the PSO algorithm has a special ability to exploit the search space, this algorithm is used in the exploitation phase to solve the problem. The accuracy obtained from our proposed algorithm (PSO-GSA) shows an improvement in the accuracy of the GSA algorithm.

Keywords: sentimental analysis (SA); particle swarm optimization (PSO); gravitational search algorithm (GSA); hybrid bio-inspired approach; heuristic search algorithms

Paper Type: Original Research

1. Introduction

Currently, social networks are a platform for online conversations where human beings make contributions to create content and share. Today, twitter is the quickest messenger among social media that is very convenient to use. The content of tweets topics such as marketing, politics, and the smallest detail in people's daily lives (Basari et al., 2013). About 200 billion tweets are made in a year, which is equivalent to 500 million daily tweets, 350,000 tweets minutely, and an average of 6,000 tweets per moment (Sayce, 2020), producing a big quantity of unstructured textual content data. The text of the generated tweets is processed and classified as data to collect the opinions of the categorized people. Sentiment analysis or data mining extracts the sentiments and opinions of people from tweets sent by users. Twitter sentiment divides tweets into positive, negative, and neutral classes (Ahuja et al., 2019). This classification can be used to help organizations and marketers in the field of evaluating their services and the quality provided to customers and get help in the proper implementation of their business strategies (Yadav and Vishwakarma, 2020). Meta-heuristic algorithms are proposed to cross the local optimal point of heuristic algorithms. Classification of large volumes of social media data using traditional methods is a very difficult, time-consuming, and low-accuracy task, therefore researchers believe that these methods can be used for efficient feature selection. Swarm intelligence-based stochastic methods are a field of artificial intelligence consisting of an organized set of particles (factors), each of which has a relatively simple structure, and the swarm social structure is a very complex relationship between group behavior and individual behavior. Individual behaviors of particles are based on their actual role in their natural habitat. The movement of agents is individual and wide. However, interactions between particles provide a collective intelligent behavior (Kumar et al., 2019).Based on past research, questions are raised, including: Given the stochastic nature of the task and the heavy dependence of sentiment analysis on the data, is it possible to propose a meta-heuristic algorithm that shows higher accuracy on all data sets? At each step of the sentiment analysis process, what techniques should we use to help improve the performance of the algorithm? What newer meta-heuristic algorithms can be suggested? And this proposal the result of the combination of previous meta-heuristic algorithms or a new algorithm inspired by nature? These are questions that researchers have been trying to answer for more than a decade.Many meta-heuristic algorithms have been used in sentiment analysis, in the meantime, the GSA algorithm, contrary to its great performance in searching the state space, has been used in a few numbers, so we used GSA in this research and examined the results. Among the completely unexplored fields of sentiment analysis are hybrid algorithms resulting from previously introduced algorithms, which can simultaneously use the advantages of both algorithms (Yadav and Vishwakarma, 2020). In this research, in addition to addressing this gap, we have proposed a hybrid PSO-GSA algorithm to solve the

[*] Corresponding Author: hosseini@kntu.ac.ir

problem raised by the researchers (Mosavi et al., 2021). PSO imitates the collective behavior of birds. In PSO, the movement of each particle in the search space is based on two factors: one is the best-known local position by that particle and the other is the best-known global position by other particles. Finally, all particles move in groups towards the best solution (Lai et al., 2009). In the year of 2009, Rashedi et al. introduced a new innovative algorithm to search for the best solution in the search space based on Newton's physical theory called GSA (Newton, 1729), which states: "Each particle in the universe absorbs another particle with a force that is directly related to the mass of two particles and the square of the distance between two particles has an inverse relationship". In this article; we use PSO, GSA algorithms, and their combination for sentiments analysis (SA). The algorithms mentioned on the data set used by Pandey et al. (2017): Testdata.manual.2009.06.14 (498 samples, 3 classes), Twitter-sanders-apple2 (479 samples, 2 classes), Twitter-sanders-apple3 (988 samples, 3 classes), Twitter dataset (2000 samples, 2 classes). In this study, we show the performance (accuracy) of algorithms in the datasets for optimal feature selection in SA. The proposed method is a new method to improve GSA performance.We will continue this paper in 4 sections: In section 2, we review the past literature. In section 3, The details of the proposed hybrid method are mentioned in this article. Section 4, includes a review of the results. Section 5, the conclusions of this study. The last part (section 6), the implications of this study for marketers.

2. Related work

The study of content and its analysis dates back to 1966. Twitter tweets have been a common source of data for analyzing positive or negative comments in past articles (Goel and Garg, 2018). In the previous series of works, different combinations of meta-heuristic algorithms and different classifications have been proposed to improve and effectively solve optimization problems. More recently, research has focused on choosing features derived from environmental life, especially swarm intelligence. Optimization is an important issue in the current research. As Yadav and Vishwakarma (2020) state in this context, nature-inspired algorithms are algorithms that help in solving optimization problems. And they find the best possible solution among a wide range of solutions. For example, traders are always trying to optimize their market decisions and strategies. Nguyen et al. (2014) presented a new approach for feature selection based on PSO and local search that mimics the conventional inverse elimination feature selection method. Their goal is to take advantage of both filtering and packaging approaches. They tested the proposed approach on eight benchmark datasets and the results showed that their proposed approach chose from three algorithms based on PSO and two traditional methods, with higher performance and fewer features. Menghour et al. (2016) proposed three different combinations of bio-inspired algorithms PSO and ant colony optimization (ACO). According to the results of their experiments, the hybrid algorithm ACO-PSO1 had the best result in terms of accuracy compared to others. Chen et al. (2019) combined the bat algorithm (BA) and the Spark parallel computing framework and proposed the SBATFS algorithm was proposed to solve the problem of the long execution time of high-dimensional data of the BA. Results show that the SBATFS algorithm was able to solve the problem and be used effectively to select the feature.The research by Astuti and Taufan (2022) has carried out sentiment analysis in two stages. In the first stage, without using meta-heuristic algorithms and using NB and SVM classification algorithms, the accuracy has been obtained. And in the next step, he used the PSO meta-heuristic algorithm. The results showed them that NB with PSO increased by 6.38 degrees compared to NB and SVM with PSO increased by 3.83 degrees compared to SVM, which indicates the efficiency and effectiveness of PSO in feature selection.Tawhid et al. (2018) have proposed a hybrid bat (BA) and PSO algorithm called HBBEPSO to effectively solve feature selection problems. The reason for using these two algorithms is that the bat algorithm has a very good ability to converge and PSO has a special ability to exploit the search space. The test results showed that the proposed HBBEPSO algorithm has a very good ability to find optimal features.Yuvaraj et al (2017) introduced a new approach to Binary Shuffled Frog Algorithm (BSFA). In this study, they used TF-IDF to extract the features and KNN, NB, and radial basis function (RBF) networks to classify. The proposed algorithm with the RBF classifier performed better than the others.Yang and Suash (2009), for the first time, proposed the CS algorithm based on cuckoo reproduction behavior to solve optimization problems. Or for the first time in 2009, Yang (2009) formulated the firefly algorithm (FA), and in his research, he examined its similarities and differences with the PSO algorithm.In 2007, Karaboga and Basturk developed an artificial bee colony (ABC) algorithm to optimize multivariable functions.In 2009, Rashedi et al. Introduced a new optimization algorithm (GSA) based on the law of gravity and mass interplay. In the proposed algorithm, the search space consists of a set of factors (masses) that interact with each other according to the laws of gravity and motion.The research of Botchway et al. (2022) used the binary mode of the PSO algorithm called BPSO to increase the accuracy of the work. The results showed that in each of the classification algorithms KNN, NB, and SVM, the accuracy increased by 0.91%, 11.6%, and 8.43%, respectively.Goel and Garg (2018) examined the GSA algorithm and compared it with the ACO algorithm. The outcome shows the superiority of the GSA algorithm over ACO. In a series of future works to develop the algorithm, he proposes combining the algorithm with other meta-heuristic algorithms.Ighazaran et al. (2018) have examined the advantages and disadvantages of feature selection (FS) in SA. The results prove the potential ability of meta-heuristic algorithms as FS in SA and they can be used to select optimal features from customer feedback. Or a review article (Yadav and Vishwakarma, 2020) in which eight bio-inspired meta-heuristic algorithms were tested under the same conditions, and the articles reviewed by it were from 2010 to 2019. The results are as follows: PSO algorithm first and then ACO has the best performance among the eight algorithms considered. Finally, according to Goel and Garg (2018) and the superiority of the PSO algorithm in most of the previous articles and according to the review done by Mosavi et al. (2021) on the quality of GSA work, in this article, we decided to combine the PSO and GSA algorithms for the first time in SA and examine the results. A number of the articles reviewed by us are given in "table 1". According to the studies conducted, due to the different performances of meta-heuristic algorithms in different data sets, the range of accuracy of the articles in the past was from 40 to 100.

Table 1. A review of the subject literature

Ref	Algorithm	Data Set	Language	Accuracy
Yousefpour et al. (2016)	GA and HS	book, electronic, music review	English	Between 80 to 95
Gokalp et al. (2020)	IG	4 amazon products and 9 5public sentiment analysis datasets	English	Up to 90
Yan et al. (2018)	BCROSAT	first nine benchmark datasets	English	Between 66 to 100
Botchway et al. (2022)	BPSO	UCI ML repository	English	Between 69 to 87
Goel and Garg (2018)	GSA	Twitter API	English	Between 73 to 93
Pandey et al. (2017)	CS and CSK	4 datasets from Twitter	English	Between 50 to 84
Nguyen et al. (2014)	PSO+Backward elimination method	UCI machine learning repository	English	Between 78 to 98
Chen et al. (2019)	SBATFS	Fudan corpus	English	Between 80 to 82
Astuti and Taufan (2022)	PSO+NB and PSO+SVM	Twitter (#Vaksin Covid-19)	Indonesian	Between 70 to 76
Tawhid et al. (2018)	HBEPSOB	UCI machine learning repository	English	Between 42 to 98
Yuvaraj et al (2017)	BSFA	Twitter corpus	English	Between 91 to 93
Alarifi et al. (2020)	CSO-LSTMNN	online marketplace Amazon	English	Between 89 to 96

3. Methodology

The initial step involves collecting the required data. In this article, four tweet datasets are considered that have a positive or negative, or neutral tag. Then, eliminate any noise, conflict, and imperfections in it so that the dataset is ready for pre-processing. The texts of the output tweets from the previous step are weighted using tf-idf to select features in the next step using meta-heuristic algorithms. The PSO, GSA algorithm, and combining both (PSO-GSA) on the property matrix were then used to obtain a set of optimal properties, then implemented in Python for training in different classifiers. "Figure 1" shows a flowchart of the work process of the present research.

We will continue Section 3 in 5 sub-sections: The data set used in this research, the pre-processing techniques used, the method of weighting the features, the meta-heuristic algorithms used, and the classification algorithm selected in this research are described.

3.1. Data collection

In this article; we tested the accuracy of the meta-heuristic algorithms PSO, GSA, and PSO-GSA on the following four twitter datasets, which are also used by Pandey et al. (2017) and contain tweets on various topics. "Table 2" provides the specifications of the datasets used.

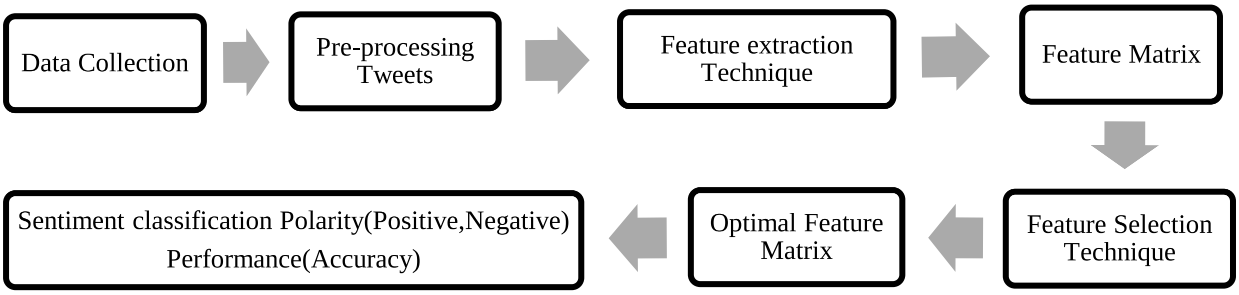

Data Collection
Pre-processing Tweets
Feature extraction Technique
Feature Matrix
Feature Selection Technique
Optimal Feature Matrix
Sentiment classification Polarity(Positive,Negative)
Performance(Accuracy)

Figure 1. The framework of the system

Table 2. Considered Twitter datasets (Chandra Pandey et al., 2017).

Sr.No	Dataset	Number of Instances	Number of classes	Posi-tive	Nega-tive	Neu-tral	Date Range	Topic Covered
1.	Testdata.manual.2009.06.14	498	3	182	177	139	May 11, 2009, to Jun 14, 2009	Google, Obama, Kindle, China, etc.
2.	Twitter-sanders-apple2	479	2	163	316	-	Oct 15, 2011, to Oct 20, 2011	Apple, Google, Microsoft, Twitter
3.	Twitter-sanders-apple3	988	3	163	316	509	Oct 15, 2011, to Oct 20, 2011	Apple, Google, Microsoft, Twitter
4.	Twitter dataset	2000	2	1000	1000	-	Nov 17, 2014, to Dec 10, 2014	sports, saints, funny Images, etc.

3.1.1. Testdata.manual.2009.06.14

This dataset contains 1.6 million tweets, divided into educational and experimental collections, from the Stanford Twitter collection from 11 May 2009 to 14 July 2009, on various subjects like China, Obama, Google, Nike, Kindle, North Korea, San Francisco, Insects, Dentists, and Iran (Astuti and Taufan). In this article; An experimental subset was used, which included 498 tweets with a subset of 182 positive tweets with a mark of "4", 177 negative tweets with a mark of "0", and 139 neutral tweets with a mark of "2".

3.1.2. Twitter-sanders-apple

Sanders Analytics compiled two datasets for Apple from 15 October 2011 to 20 October 2011, based on various topics: Twitter, Google, Apple, and Microsoft. Each tweet was manually tagged by Nick Sanders as "pos", "neg" or "neutral".

3.1.2.1. Twitter-sanders-apple2

This dataset (Botchway et al.) is a subset of the Twitter-sanders-apple and contains 163 positive tweets and 316 negative tweets for a total of 479 tweets.

3.1.2.2. Twitter-sanders-apple3

Twitter-sanders-apple3 is another subset of Twitter-sanders-apple (Botchway et al.). and has three classes with 163 positive tweets, 316 negatives, and 509 neutral tweets for a total of 988 tweets.

3.1.3. Twitter dataset

This dataset ("Twitter dataset," 2014) is taken from twitter based on the topics of saints, sports, jokes, students, and funny pictures. This dataset was collected from November 17, 2014, to December 10, 2014, including 2,000 tweets. The dataset is manually labeled 1000 positive tweets with a mark of 1 and 1000 negative tweets with a mark of 0.

3.2. Pre-Processing

Symeonidis et al. (2018) believed that the right combination of pre-processing techniques could increase the accuracy of classification. Therefore, according to "Table 3", Symeonidis selected 17 commonly used techniques in pre-processing and tested their performance under the same conditions. According to the results, the combination of replacing contractions, replacing repetitions of punctuation, replacing URLs and user mentions, removing numbers, and lemmatizing is the best combination of preprocessing in the field of sentiment analysis.

Jianqiang and Xiaolin (2017) believed that deleting stop words, numbers, and URL reduces noise but does not affect accuracy performance. Replacing negation and expanding words (abbreviations) is effective in sentiment analysis.

Due to the nature of the data set considered in this article and the set of results from previous studies, we considered the pre-processing combination that provides better results in classification:

- **Substitute usernames and URLs:** In Twitter texts, most sentences contain a username, URL, or hashtag symbol that does not evoke any sentiment. We eliminated these in our work.
- **Replacement of contractions**: Strings such as "haven't " and "didn’t" will be replaced by "have not" and "did not," respectively. If the contractions are not replaced, the token process creates the tokens "didn" and "'t" (for the case of "didn’t"), which are two meaningless tokens and may be deleted later from the tweet. In our work, we replaced contractions with their complete forms.
-

Table 3. Correspondence of pre-processing techniques (Symeonidis et al., 2018).

Number	Pre-processing Technique
0	Basic (Remove Unicode strings and noise)
1	Other (Replace URLs and user mentions)
2	Replace Slang and Abbreviations
3	Replace Contractions
4	Remove Numbers
5	Replace Repetitions of Punctuation
6	Replace Negations with Antonyms
7	Remove Punctuation
8	Handling Capitalized Words
9	Lowercase
10	Remove Stop words
11	Replace Elongated Words
12	Spelling Correction
13	Part of Speech Tagging
14	Lemmatizing
15	Stemming
16	Handling Negations

- **Emoji:** In some of the considered tweets, there is an emoji ASCII code, which we replaced with the equivalent emotional words.
- **Replacing slang and abbreviations:** Users on social media usually write informally and their texts contain many slangs and abbreviations. Slang is a language that consists of words and phrases that are considered very informal. In short, an abbreviation or abbreviated form is a word or phrase. For a correct interpretation, these words and phrases must be replaced to convey their meaning correctly (Jianqiang and Xiaolin, 2017).
- **Replacing repetitions of punctuation:** The writing symbols that indicate emotions are: stop signs, question marks, and exclamation marks. The repeated use of these punctuation marks indicates strong emotions, which we replace with a delegate label. For example, the "???" sign Replace with "multiQuestionMark."
- **Removing numbers:** Numbers do not contain any emotion and removing them from the text is one of the most used pre-processing methods. Some slang words like "W8" which means "wait…" contain numbers, so this step must be done after replacing slang (Jianqiang and Xiaolin, 2017).
- **Lowercasing:** By doing this, the words are uniform in appearance, and the problems are reduced.
- **Spelling correction:** Spelling mistakes by users in informal texts are so common that they can make classification difficult and reduce accuracy. For this purpose, words were matched with a dictionary.
- **Stemming:** Rooting is returning words to their root forms by removing prefixes and suffixes. This technique merges many words and reduces the data size (Jianqiang and Xiaolin, 2017).

Replacing elongated words: Elongated is a word that one character is mistakenly (but often purposely) repeated one or more times, for example, "greeeeat". So, such words are replaced with their source word to be integrated. Otherwise, in the classification stage, a unique word is identified and deleted due to its low frequency (Jianqiang and Xiaolin, 2017).

3.3. Feature extraction (TF-IDF)

In the final part of pre-processing, all texts are weighted using the tf-idf technique. We used one of the most common word weighting methods, tf-idf, which is used in various subject areas, including text categorization and summarization. TF-IDF is a numeric statistic that shows the importance (weight) of each word in a tweet (document) relative to all tweets (data set).

TF-IDF is a statistical technique that determines the importance of a word in a document based on the number of times it appears in that document and a specific set of documents (dataset). The term frequency of a particular phrase is calculated as the number of times a phrase in a tweet equals the total number of words in a tweet. For example, if a word is repeated more often than others, it means that it is a more important and relevant word than the others and it is assigned a high score (TF). IDF (Reverse Document Frequency) means how much information a word provides about the document in which it appears. There are some terms like "of", "the", "a" etc., but they do not give much information. IDF is calculated as IDF (t) = log10 (N / DF), where N is the number of tweets and DF is the number of tweets including the expression t. Suppose there is a tweet that includes 400 words and out of these 400 words the "home" appears 20 times, from this case the frequency of the phrase will be 20/400 = 0.05, and suppose there are 100000 tweets, and out of this number, only 1000 tweets include home so IDF (home) = log10(100000/1000) = 2, and TF-IDF (home) 0.05 * 2 = 0.1 (Ahuja et al., 2019).

3.4. Algorithms

The general purpose of this article is to compare the performance of the GSA algorithm with the PSO algorithm under the same conditions. According to (Mosavi et al., 2021; Zhou et al., 2015) which is in areas other than sentiment analysis, the GSA algorithm has an excellent ability to search overall, but in the last iterations, it has a low speed in exploiting the search space. Due to the unique ability of particle swarm optimization (PSO) in the operation phase, this method is used to solve the above problem.

3.4.1. Particle swarm optimization (PSO)

PSO (Kennedy and Eberhart, 1995) is based on swarm intelligence, inspired by the social behavior of birds. In PSO, particles are candidate solutions that move en masse to reach the global optimum. During motion, each particle (i) has a position and velocity that indicates xi = {xi1, xi2, …, xiD} and vi = {vi1, vi2, ..., viD}, respectively. Where D indicates the dimensions of the search space. Each particle can remember its best position (pbest) ever visited, and the best previous position ever visited by the whole group (birds), called the best global position (gbest). In each iteration, based on pbest and gbest, xi and vi are updated for each particle to search for optimal solutions according to the equations. 1 and 2.

$$v_{id}^{t+1} = w * v_{id}^{t} + c_1 * r_{i1} * (p_{id} - x_{id}^{t}) + c_2 * r_{i2} * (p_{gd} - x_{id}^{t}) \quad (1)$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \quad (2)$$

where v_{id}^{t+1} shows the velocity of particle i in the dth dimension at the (t + 1) th iteration. w is the inertia weight reflecting the influence of the previous velocity. ri1 and ri2 are random values, which are uniformly distributed in [0, 1]. c1 and c2 are acceleration constants. x_{id}^{t+1}shows the position value of particle i. pgd and pid shows the values of gbest and pbest in the dth dimension (Nguyen et al., 2014).

3.4.2. Gravitational search algorithm (GSA)

The GSA algorithm (Rashedi et al., 2009) is inspired by Newton's law and the iteration of masses of global gravitation. According to (Rashedi et al., 2009; Zhou et al., 2015) suppose there is a system with n factor (masses). For n factor, each factor's position is:

$$x_i=\{x_i^1,\ldots,x_i^d,\ldots x_i^n\}, \qquad \text{for i=1, 2, …, N} \quad (3)$$

Where xdi in the dth dimension presents the position of ith factor.

All factors are randomly initialized in an initial space (problem). In repetition t, the power of mass j on mass i is defined as follows:

$$\mathrm{F}_{\mathrm{ij}}^{\mathrm{d}}(\mathrm{t}) = \mathrm{G}(\mathrm{t}) \frac{\mathrm{M}_{\mathrm{pi}}(\mathrm{t}) \times \mathrm{M}_{\mathrm{aj}}(\mathrm{t})}{\mathrm{R}_{\mathrm{ij}}(\mathrm{t}) + \varepsilon} (\mathrm{x}_{\mathrm{j}}^{\mathrm{d}}(\mathrm{t}) - \mathrm{x}_{\mathrm{i}}^{\mathrm{d}}(\mathrm{t})) \tag{4}$$

where G(t) is the gravitational constant at time t, Mpi is the passive gravitational mass related to factor i, Maj is the active gravitational mass related to factor j, ε is a small constant, and Rij(t) is the euclidian distance between two factors i and j:

$$R_{ij}(t) = \| x_i(t). x_j(t) \| 2 \tag{5}$$

The following formula calculates the total gravitational force of factor i in the search space:

$$\mathrm{F}_i(t) = \sum_{j=1. j \neq i}^{n} rand_j \times F_{ij}(t) \tag{6}$$

where randj is a random number in the interval [0, 1].

so, the accelerations of all factors are calculated as follows:

$$a_i(t) = \frac{\mathrm{F}_i(t)}{M_{ii}(t)} \tag{7}$$

where Mii is the inertial mass of ith factor.

The velocity and position of the factors are defined as follows:

$$v_i(t+1) = rand_i \times v_i(t) + a_i(t) \tag{8}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{9}$$

where randi is a uniform random variable in the interval [0, 1]. This is a random number to have a random variable in the search process.

The masses of all factors are updated using the following equation:

$$\mathrm{M}_i(t) = \frac{fit_i(t) - worst_i(t)}{best_i(t) - worst_i(t)} \tag{10}$$

The best or the worst depends on the type of problem being studied. That is the problem of maximizing or the problem of minimizing.

GSA considers the interaction between particles, but due to a lack of memory, agents cannot see the optimal world. The main drawback of the algorithm is that its convergence speed is slow.

3.4.3. Hybrid PSO and GSA (PSO-GSA)

PSO-GSA (Figure 2) is the approach proposed by Menghour and Souici-Meslati (2016) where our idea is to replace ant colony optimization algorithms (ACO) with GSA.
As mentioned earlier, GSA has an excellent ability to search overall, but in the last iterations, it has a low speed in exploiting the search space. Due to the special capability of PSO in the operation phase, this method is used to resolve the above problem.
According to Rashedi et al. (2009), Some critical differences between the two algorithms are as follows:

- In PSO, the direction of a factor is determined depending on the two best positions, pbesti and gbest. But in GSA the direction of the factor is calculated based on the resultant force of other factors.
- In PSO, updates are performed regardless of the quality of the solutions and the appropriateness values. At the same time, in GSA, the force is calculated in proportion to the fitness value. Therefore, factors in the search space find themselves under the influence of force.
- PSO uses memory to update speed (due to pbesti and gbest). However, due to the lack of memory, the GSA is updated depending on the current location of the factors.
- In PSO, the distance between the solutions does not make sense, while in GSA, the force is inversely proportional to the distance between the solutions.
- Ultimately, the idea of PSO search is inspired by the social behavior of birds, while GSA is inspired by a physical phenomenon.

We used the advantages of both algorithms in our proposed algorithm (PSO-GSA) to improve the results of the GSA algorithm.

3.5. Classification algorithm

Shojaei et al. (2021) acknowledged that today, classification problems have been considered by many researchers. In these problems, objects are classified into different classes according to their similarities or differences. For example, an object is considered a property vector, and depending on its properties, it is placed in its respective class, and objects are classified.
In our work, we selected six classifiers so that we can compare the results and choose a better classifier.

3.5.1. K-Nearest neighbour (KNN)

It is a simple algorithm commonly used in intrusion detection, pattern detection, and other cases. The method of the algorithm is that the Euclidean distance (other criteria such as Manhattan distance, etc.) compares the new data point with the data points in the classes and considers the closest distance to the value of k as the new data point class (Imandoust and Bolandraftar, 2013).

3.5.2. Naïve bayesian classifier (NB)

It is a powerful classification algorithm for large and small data dimensions. Among its advantages are its speed and scalability, which can be used in binary and multi-class classifications. It works on a probability basis and works on Bayes' theorem. In the Naïve Bayes class, the predicted class is considered the most probable. Naïve Bayes, additionally recognized as Maximum a Posterior Naïve Bayes, has various advantages and disadvantages in various scientific fields (Rathi et al., 2018).

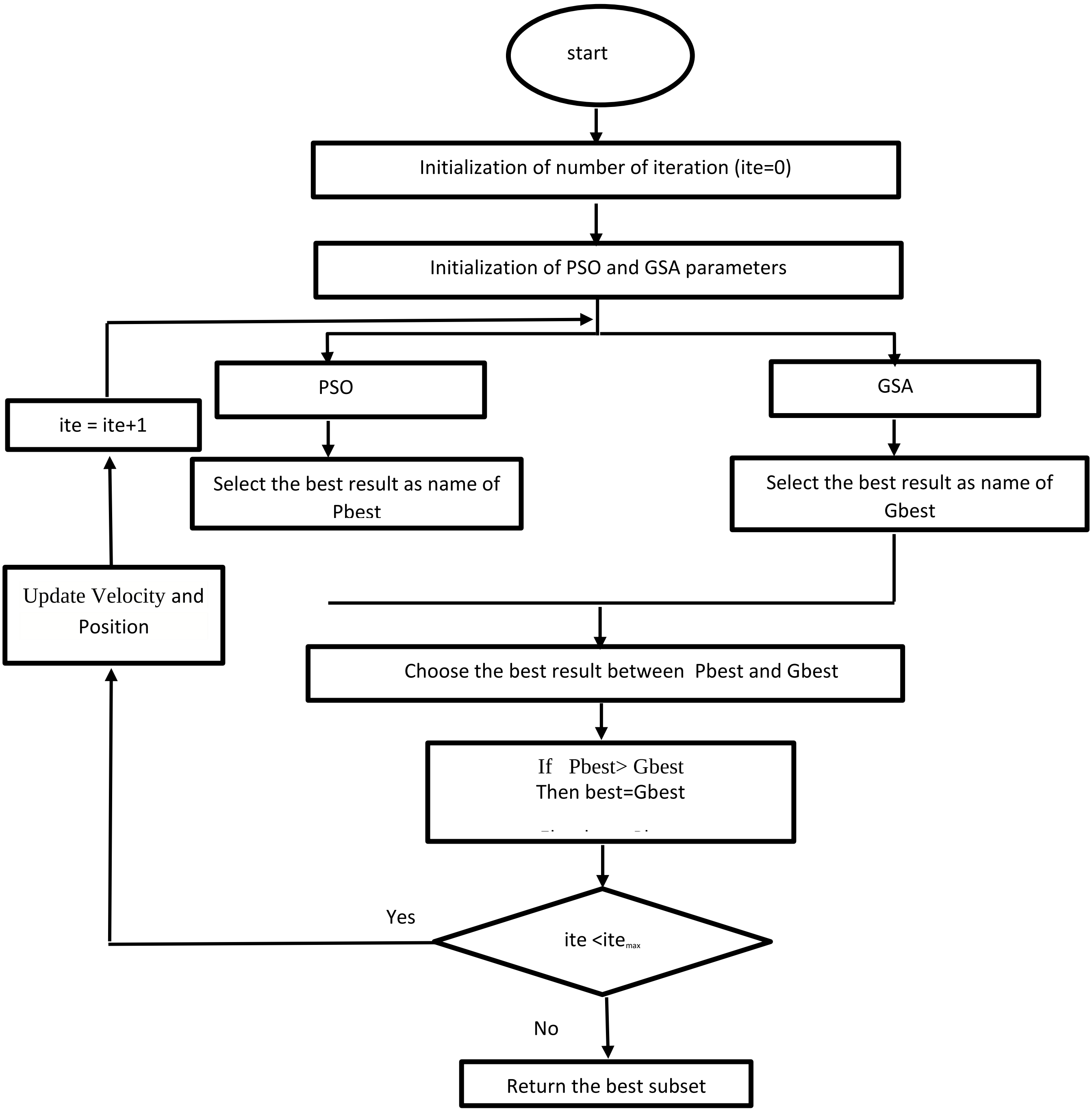

start
Initialization of number of iteration (ite=0)
Initialization of PSO and GSA parameters
PSO
GSA
ite = ite+1
Select the best result as name of Pbest
Select the best result as name of Gbest
Update Velocity and Position
Choose the best result between Pbest and Gbest
If Pbest> Gbest
Then best=Gbest
Yes
ite <ite$_{max}$
No
Return the best subset

Figure 2. PSO-GSA hybrid approach

3.5.3. Decision tree (DT)

This algorithm is used in the fields of regression and classification and management of categorical and numerical data. The way it works is that it divides the data set into smaller subsets, and at the same moment, the relational tree is created (Acharya et al., 2012).

3.5.4. Multilayer perceptron (MLP)

ANNs are computational models that are evoked by the biological model of the brain (Ramchoun et al., 2016). MLPs are a fully connected class of feed-forward ANN. It consists of at least three inputs, hidden, and output layers. Except for input nodes, each node is identified by a neuron that uses a nonlinear activation function. Supervised learning in MLP is a post-diffusion technique. Its layered nature adds to its ability to differentiate nonlinear data.

3.5.5. Support vector machine (SVM)

This is an unlikely binary linear classification algorithm used for regression as well as classification purposes. The SVM draws training samples (points) in space so that they become two categories, then draws new samples (new points) based on the prediction of belonging to which category in the same space. It performs well in regression and high-dimensional data because the SVM effect increases with increasing space dimensions (İ et al., 2017).

3.5.6. Logistic regression

Is a type of statistical model used for classification and analysis. Logistic regression estimates the probability of an event occurring based on a data set of independent variables. Since the result is a probability, the dependent variable is limited between 0 and 1. This algorithm belongs to the generalized linear model class (Pedregosa et al., 2011).

4. Experimental results

The accuracy and efficiency of the three methods of PSO, GSA algorithm, and combining both (PSO-GSA) have been tested on the four twitter datasets (Table 1), which contain tweets on various topics.In the first step, we tested PSO with different classifications on the data so that we could find the best classification compared to our data set. In the next step, we tested the three algorithms with the superior classification of the previous step. According to "Figure 3", the K-Nearest Neighbors (KNN) algorithm has shown higher accuracy than other classification algorithms in all four datasets so we choose the KNN algorithm as a classifier and continue working with it.

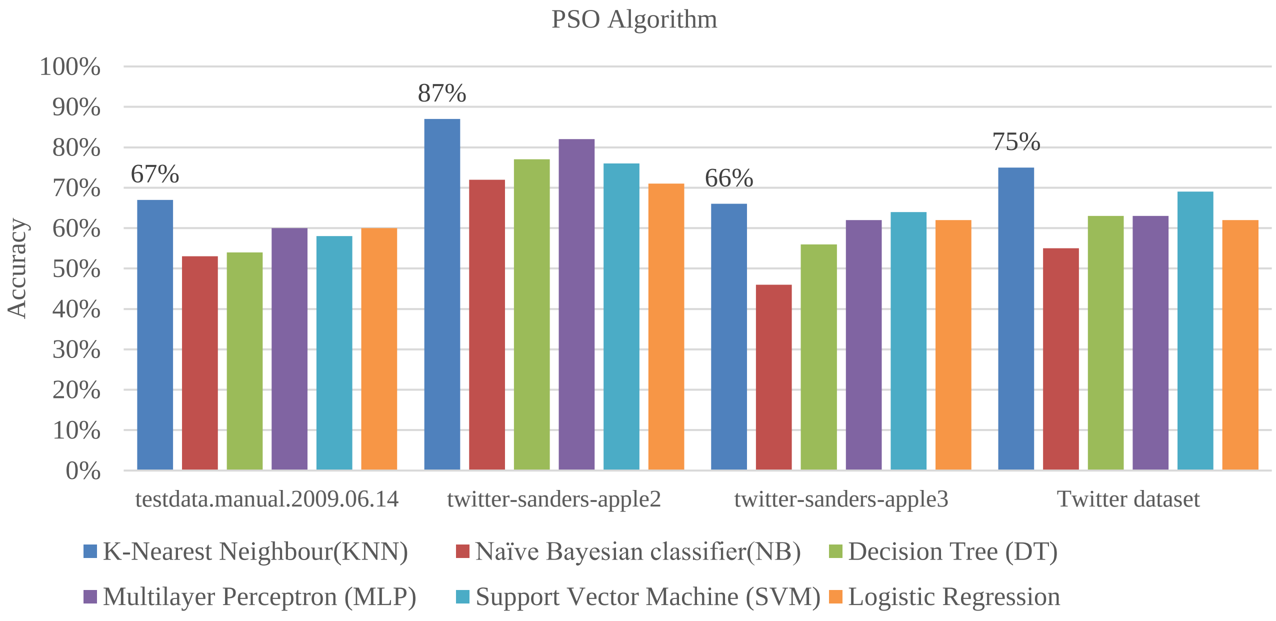


Figure 3. Comparison of The Performance of Different Classification on Four Datasets

We have divided each of the four datasets into educational sections for educational and experimental purposes with a split ratio of 0.8 (80% training data and 20% test data). On each of the 4 data sets, three algorithms, PSO, GSA algorithm, and combining both (PSO-GSA) with the number of iterations (T = 1000) and population size (n = 10) have been implemented in 30 steps. Finally, the average accuracy is performed in 30 steps and is shown in "Table 4". Also, the results of three different criteria precision, recall, and F1-score are given in "Table 5".We set the parameters of PSO as follows. C1 and C2=2, and w is set to 0.8, while for GSA parameters, we set: G0 =1 and alpha=20.Pandey et al. (2017), test the PSO algorithm with the k-means classifier. The results of the four selected datasets shared with this article are shown in "Table 6". According to the report of this article and the results of our experiment for the PSO algorithm, it can be concluded that we were able to improve the results of PSO by changing the classification algorithm from K-means to KNN.Yadav and Vishwakarma (2020) reported the accuracy superiority of the PSO algorithm among nine bio-inspired meta-heuristic algorithms.

Table 4. Accuracy Obtained Using Optimized Approaches.

Sr.No	Datasets	Method	Mean Accuracy
1	testdata. manual.2009.06.14	PSO	67%
		GSA	64%
		PSO-GSA	68%
2	twitter-sanders-apple2	PSO	87%
		GSA	81%
		PSO-GSA	82%
3	twitter-sanders-apple3	PSO	66%
		GSA	62%
		PSO-GSA	63%
4	twitter dataset	PSO	75%
		GSA	65%
		PSO-GSA	67%

Table 5. Computational results of three metaheuristic algorithms on four Twitter datasets.

Dataset	Method	Positive			Negative			Neutral		
		Precision	Recall	F1score	Precision	Recall	F1score	Precision	Recall	F1score
testdata. manual.2009.06.14	PSO	0.72	0.72	0.70	0.67	0.63	0.67	0.67	0.67	0.62
	GSA	0.67	0.64	0.65	0.68	0.69	0.68	0.62	0.56	0.57
	PSO-GSA	0.70	0.71	0.72	0.70	0.72	0.68	0.68	0.57	0.60
twitter-sanders-apple2	PSO	0.88	0.74	0.80	0.87	0.945	0.90	-	-	-
	GSA	0.77	0.44	0.51	0.75	0.87	0.85	-	-	-
	PSO-GSA	0.78	0.59	0.68	0.81	0.92	0.87	-	-	-
twitter-sanders-apple3	PSO	0.69	0.29	0.39	0.71	0.49	0.58	0.63	0.89	0.75
	GSA	0.57	0.23	0.33	0.62	0.43	0.52	0.62	0.81	0.71
	PSO-GSA	0.66	0.27	0.37	0.67	0.47	0.53	0.63	0.89	0.71
twitter dataset	PSO	0.77	0.70	0.75	0.74	0.78	0.73	-	-	-
	GSA	0.69	0.63	0.65	0.66	0.70	0.66	-	-	-
	PSO-GSA	0.69	0.65	0.67	0.67	0.71	0.68	-	-	-

Table 6. Accuracy Obtained Using Optimized Approach.

Sr.No	Datasets	Method	Mean Accuracy with K-means [8]	Mean Accuracy (Current Research) with KNN
1	testdata. manual.2009.06.14	PSO	59.28%	67%
2	twitter-sanders-apple2	PSO	57.24%	87%
3	twitter-sanders-apple3	PSO	62.17%	66%
4	Twitter dataset	PSO	50.55%	75%

The GSA algorithm was not among the nine algorithms considered by Yadav and Vishwakarma (2020). According to the reported results, among the nine bio-inspired algorithms, the PSO algorithm with the SVM classifier performed best in terms of accuracy. So, we used the GSA algorithm with the KNN classifier to compare with the PSO. Based on the results of this comparison ("Figure 4"), we found that PSO still performed better than GSA in all datasets.Articles (Mosavi et al., 2021; Zhou et al., 2015) stated that GSA has an excellent ability to search overall, but in the last iterations, it has a low speed in exploiting the search space. Since the PSO algorithm has a special ability to exploit the search space, this algorithm is used in the exploitation phase to solve the problem. For this purpose, we got ideas from these articles and used a combination of two algorithms for sentiment analysis.

The results show that the PSO-GSA algorithm has increased the performance (accuracy) of GSA. For example, the testdata. manual.2009.06.14 dataset has improved from an average accuracy of 64% in the GSA algorithm to 68% in the PSO-GSA algorithm. Also, in all datasets except testdata. manual.2009.06.14, PSO still has the highest level of accuracy.

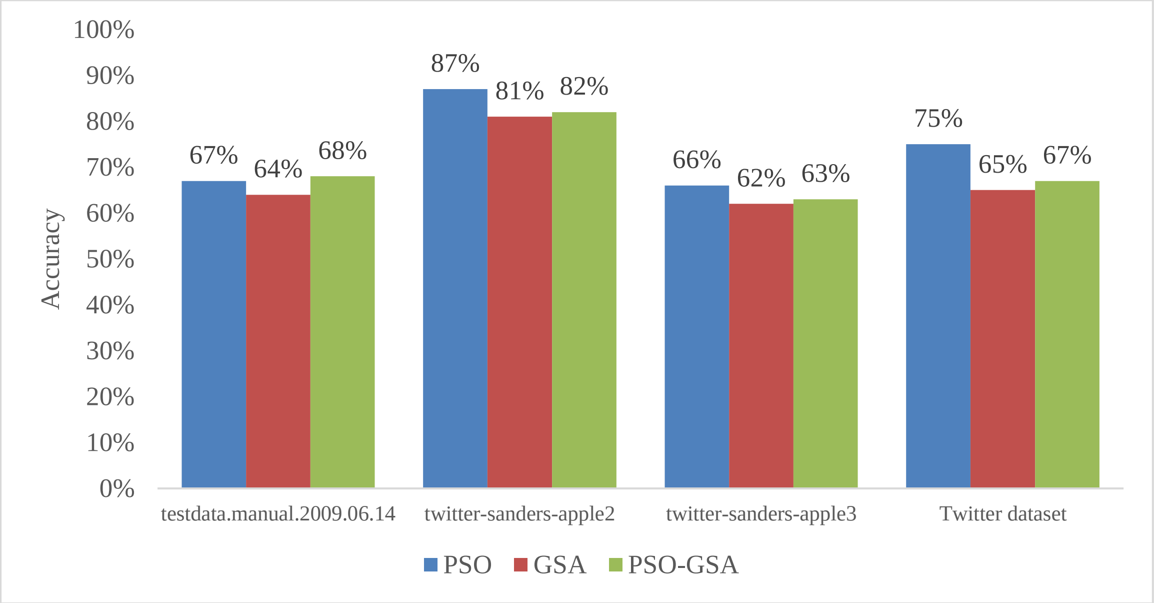


Figure 4. Accuracy Obtained Using Optimized Approaches of Current Research.

5. Conclusion

In this research, meta-heuristic algorithms for feature selection (FS) were tested for the first time in sentiment analysis, we proposed an algorithm that is a combination of particle swarm optimization (PSO) and gravity search (GSA) algorithms. Which was tested on four databases. In this research, our first hypothesis to improve the performance of GSA was to combine it with PSO, which was proven according to the results obtained in Table 4 and Table 5. In the next hypothesis, we expected that the proposed hybrid algorithm could perform better than both GSA and PSO, but the PSO-GSA algorithm could not perform better than PSO. In the third hypothesis, we had predicted that the PSO algorithm used in this research would show a better performance than the similar research, which according to Table 6, this prediction was made correctly. It can be said that in our evaluation section, the performance of the proposed algorithm was compared with the standard model of PSO and GSA algorithms. Our experimental studies showed that the PSO-GSA algorithm effectively improves the accuracy and selects fewer features. PSO-GSA was able to improve the classification accuracy of the GSA algorithm and according to the results obtained in this article, the proposed algorithm in the database test data. manual.2009.06.14 was able to achieve higher accuracy than PSO. In the future research series, we plan to use the divergence feature of GSA in the early convergence of PSO to improve the accuracy of PSO-GSA compared to PSO. More generally, we can combine the GSA algorithm with other meta-heuristic algorithms such as genetics, cuckoos, etc. to improve its performance.

6. Implication

The development of technology and the increase in popularity of social media among people has provided a new opportunity for marketers to measure and analyze the quality of their services using the psychology of consumer emotions and social media. By using sentiment analysis, marketers can effectively and efficiently evaluate the opinions of their customers about the services provided and use them to increase the sales of their products. The gender of customers does not affect sentiment analysis. However, the location of customers is influential because customers in states different from the target location give more negative opinions, which indicates that cultural and social influences should also be considered in sentiment analysis. For this reason, marketers can understand and plan unique marketing strategies for each target market by using sentiment analysis according to the location and emotional location of their customers to increase the number of customers and achieve multiple profitability.

In this research, we were able to analyze user sentiments in target tweets and increase accuracy and reduce time compared to traditional algorithms.

References

Acharya, U. R., Molinari, F., Sree, S. V., Chattopadhyay, S., Ng, K.-H. and Suri, J. S. (2012). Automated diagnosis of epileptic EEG using entropies. Biomedical Signal Processing and Control, 7(4), 401-408. doi:https://doi.org/10.1016/j.bspc.2011.07.007

Ahuja, R., Chug, A., Kohli, S., Gupta, S. and Ahuja, P. (2019). The Impact of Features Extraction on the Sentiment Analysis. Procedia Computer Science, 152, 341-348. doi:https://doi.org/10.1016/j.procs.2019.05.008

Alarifi, A., Tolba, A., Al-Makhadmeh, Z. and Said, W. (2020). A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. The Journal of Supercomputing, 76(6), 4414-4429.

Astuti, F. and Taufan, R. (2022). Sentiment Analysis of Covid-19 Vaccination on Twitter Using Classification Algorithms based on PSO. Sistemasi: Jurnal Sistem Informasi, 11(2), 364-376.

Basari, A. S. H., Hussin, B., Ananta, I. G. P. and Zeniarja, J. (2013). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering, 53, 453-462. doi:https://doi.org/10.1016/j.proeng.2013.02.059

Botchway, R. K., Yadav, V., Komínková, Z. O. and Senkerik, R. (2022). Text-based feature selection using binary particle swarm optimization for sentiment analysis. Paper presented at the 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET).

Chandra Pandey, A., Singh Rajpoot, D. and Saraswat, M. (2017). Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management, 53(4), 764-779. doi:https://doi.org/10.1016/j.ipm.2017.02.004

Chen, H., Hou, Q., Han, L., Hu, Z., Ye, Z., Zeng, J. and Yuan, J. (2019, 18-21 Sept. 2019). Distributed Text Feature Selection Based On Bat Algorithm Optimization. Paper presented at the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS).

Goel, L. and Garg, A. (2018). Sentiment Analysis of Social Networking Websites using Gravitational Search Optimization Algorithm. International Journal of Applied Evolutionary Computation (IJAEC), 9(1), 76-85. doi:10.4018/IJAEC.2018010105

Gokalp, O., Tasci, E. and Ugur, A. (2020). A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification. Expert Systems with Applications, 146, 113176.

İ, İ., Atasoy, Ö. F. and Alçiçek, H. (2017, 5-8 Oct. 2017). Sentiment classification of social media data for telecommunication companies in Turkey. Paper presented at the 2017 International Conference on Computer Science and Engineering (UBMK).

Ighazran, H., Alaoui, L. and Boujiha, T. (2018, 21-23 Nov. 2018). Metaheuristic and Evolutionary Methods for Feature Selection in Sentiment Analysis (a Comparative Study). Paper presented at the 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT).

Imandoust, S. B. and Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International journal of engineering research and applications, 3(5), 605-610.

Jianqiang, Z. and Xiaolin, G. (2017). Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. IEEE Access, 5, 2870-2879. doi:10.1109/ACCESS.2017.2672677

Karaboga, D. and Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of Global Optimization, 39(3), 459-471. doi:10.1007/s10898-007-9149-x

Kennedy, J. and Eberhart, R. (1995, 27 Nov.-1 Dec. 1995). Particle swarm optimization. Paper presented at the Proceedings of ICNN'95 - International Conference on Neural Networks.

Kumar, A., Jaiswal, A., Garg, S., Verma, S. and Kumar, S. (2019). Sentiment Analysis Using Cuckoo Search for Optimized Feature Selection on Kaggle Tweets. International Journal of Information Retrieval Research (IJIRR), 9(1), 1-15. doi:10.4018/IJIRR.2019010101

Lai, C.-C., Wu, C.-H. and Tsai, M.-C. (2009). Feature selection using particle swarm optimization with application in spam filtering. International Journal of Innovative Computing, Information and Control, 5(2), 423-432.

Menghour, K. and Souici-Meslati, L. (2016). Hybrid ACO-PSO based approaches for feature selection. Int J Intell Eng Syst, 9(3), 65-79.

Mosavi, M. R., Khishe, M. and Moridi, A. (2021). Classification of Sonar Target using Hybrid Particle Swarm and Gravitational Search. Iranian journal of Marine technology, 3(1), 1-13.

Newton, I. (1729). In experimental philosophy particular propositions are inferred from the phenomena and afterwards rendered general by induction. Principia', Book, 3.

Nguyen, H. B., Xue, B., Liu, I. and Zhang, M. (2014, 6-11 July 2014). Filter based backward elimination in wrapper based PSO for feature selection in classification. Paper presented at the 2014 IEEE Congress on Evolutionary Computation (CEC).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Ramchoun, H., Ghanou, Y., Ettaouil, M. and Janati Idrissi, M. A. (2016). Multilayer perceptron: Architecture optimization and training.

Rashedi, E., Nezamabadi-pour, H. and Saryazdi, S. (2009). GSA: A Gravitational Search Algorithm. Information Sciences, 179(13), 2232-2248. doi:https://doi.org/10.1016/j.ins.2009.03.004

Rathi, M., Malik, A., Varshney, D., Sharma, R. and Mendiratta, S. (2018, 2-4 Aug. 2018). Sentiment Analysis of Tweets Using Machine Learning Approach. Paper presented at the 2018 Eleventh International Conference on Contemporary Computing (IC3).

Sayce, D. (2020). The number of tweets per day in 2020. Retrieved on October.

Shojaee, Z., Shahzadeh Fazeli, S. A., Abbasi, E. and Adibnia, F. (2021). Feature Selection based on Particle Swarm Optimization and Mutual Information. Journal of AI and Data Mining, 9(1), 39-44. doi:10.22044/jadm.2020.8857.2020

Symeonidis, S., Effrosynidis, D. and Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications, 110, 298-310. doi:https://doi.org/10.1016/j.eswa.2018.06.022

Tawhid, M. A. and Dsouza, K. B. (2018). Hybrid binary bat enhanced particle swarm optimization algorithm for solving feature selection problems. Applied Computing and Informatics.

Twitter dataset. (2014). Retrieved from https://drive.google.com/file/d/0BwPSGZHAP _ yoN2pZcVl1Qmp1OEU/view?usp=sharing

Yadav, A. and Vishwakarma, D. K. (2020). A comparative study on bio-inspired algorithms for sentiment analysis. Cluster Computing, 23(4), 2969-2989. doi:10.1007/s10586-020-03062-w

Yan, C., Ma, J., Luo, H. and Patel, A. (2019). Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. Chemometrics and Intelligent Laboratory Systems, 184, 102-111.

Yang, X.-S. (2009, 2009//). Firefly Algorithms for Multimodal Optimization. Paper presented at the Stochastic Algorithms: Foundations and Applications, Berlin, Heidelberg.

Yang, X. S. and Suash, D. (2009, 9-11 Dec. 2009). Cuckoo Search via Lévy flights. Paper presented at the 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC).

Yousefpour, A., Ibrahim, R., Hamed, H. N. A. and Yokoi, T. (2016). Integrated feature selection methods using metaheuristic algorithms for sentiment analysis. Paper presented at the Asian Conference on Intelligent Information and Database Systems.

Yuvaraj, N. and Sabari, A. (2017). Twitter Sentiment Classification Using Binary Shuffled Frog Algorithm. Intelligent Automation & Soft Computing, 23(2), 373-381. doi:10.1080/10798587.2016.1231479

Zhou, Z., Zhang, D., Sun, Z. and Wang, J. (2015, 2015//). An Adaptive Hybrid PSO and GSA Algorithm for Association Rules Mining. Paper presented at the Cloud Computing and Security, Cham.