A Novel Hybrid Algorithm to Classify Spam Profiles in Twitter

R. Krithiga
Assistant Professor, Department of Computer Science, Perunthalaivar Kamarajar Arts College, Puducherry, India.
E-mail: kriithiga@gmail.com

Dr.E. Ilavarasan
Professor, Department of Computer Science & Engg., Pondicherry Engineering College, Puducherry, India.
E-mail: eilavarasan@pec.edu

Received March 05, 2020; Accepted May 08, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V17I1/WEB17003

Abstract

Spam profile detection problem is a huge threat to social networks and poses severe challenges to safeguard the identity of users. The consequences due to the existence of spam profiles are alarming. Though several techniques have been proposed in history to identify the spam profiles, the usability of these methods is very limited due to the evolving nature of spammers. A method devised for a spammer strategy may not work when the spammers change their identity and behavior. Hence, there is a need to develop spam detector systems that are robust and work effectively even when the spammers' strategies are evolving. In this paper, a unique hybrid wrapper based technique to detect spam profiles in the online social network, MWOA-SPD is proposed. The Whale Optimization Algorithm (WOA) is integrated with the Salp Swarm Algorithm (SSA) to achieve better classification accuracy with a minimal subset of features. The exploration technique of WOA is replaced with the position updating mechanism of SSA to diversify the search and to get rid of the limitation of WOA. A dataset was extracted from Twitter and used as a benchmark to evaluate the performance of the proposed method. The findings of the results show that the proposed method yields competitive results compared to the existing ones.

Keywords

Hybrid WAO, Spam Profile Detection, Salp Swarm Algorithm, Twitter, Social Networks.

Introduction

In recent years, the online social networks (OSN) have exponential growth and the field is booming as it aids in the promotion of brands, sharing of information, staying connected

with a circle of a network that consists of family and friends. There exist several OSN's like Twitter, Facebook, Instagram, Tumblr, Sina Weibo, LinkedIn, Pinterest, etc. and depending on the services offered by each of these unique networks, the users create accounts with them. According to statistics prepared by Oberlo [1], the OSNs have 3.2 billion daily active users which roughly estimate to about 42% of the total population. This study indicates the interest shown by the people towards social media. The reason people throng the OSNs are many; following a fan/celebrity, promotion of brands, promotion of services/products, sharing of knowledge, to be a part of community groups, be in a network of family and friends to share/receive updates. Hence, the OSN has something or other to offer people of all age groups, and using them has almost become a part of our daily activities. This has attracted people around the world to show active participation in social media.

Though OSNs offer various services and benefits, it also has several threats associated with it that emanate uninvitedly. Spamming is one of the serious threats of social media and is effected by sending unsolicited messages to users to invoke malicious activities such as advertising, fetching users' data, spreading malicious content, phishing, hijacking, etc., Before the advent of social media, emails were used as the primary communication tool. As the number of email-users increased, the spammers intruded to execute their evil-activities by exploiting the network. As the email systems emerged as more sophisticated ones that recognize the spam emails accurately and trap them, the attention of spammers has now turned to social media and become the target to perform unsolicited or undesired activities. Hence, there is an urgent need to identify the spammers to protect the network.
The spam profile detection is a binary classification problem that appropriately labels the spam and legitimate profiles. During this process, misclassifying spam as a non-spam would lead to hazardous consequences, and similarly labeling a non-spammer as spam would disable the user to benefit from the network. This paper addresses the problem of spam profile identification using effective methods. The existing spam identification techniques are inadequate to catch the spammers as they keep changing their identity and exhibit various properties from time to time [2]. In this work, we propose a wrapper based hybrid technique that integrates the Whale Optimization Algorithm (WOA) with the Salp Swarm Algorithm (SSA) to solve the spam profile detection problem by embedding the operators of SSA in the search procedure of WOA.

The hybridization of meta-heuristic algorithms is expanding as a new research domain as it is developed considering the problem rather than the algorithm itself [14]. The hybrid

algorithms have also been proven to yield good results with effective run time. This has motivated us to propose a hybrid technique that would reduce misclassification errors.
The major contributions of this work are four-fold and are as follows:

- Propose a hybrid technique MWOA-SPD embedding the operators of SSA into the exploration phase of WOA
- Design a new set of robust features to manipulate the evolving spammers
- Application of the proposed method to a twitter dataset
- Analyzing the performance of the proposed technique by comparing it with neural network classifiers such as CNN, DNN, RNN, and the traditional WOA.

The remaining of the paper is structured as follows: Section II briefs the related work in the literature. Section III provides the basics of WOA and SSA along with the corresponding position updating mechanisms. Section IV presents the proposed hybrid WOA-SSA algorithm for spam profile detection. Section V showcases the experimental results and analysis of the proposed method along with the algorithms considered for the comparison. Section VI concludes the work with further insights on the scope of improvement and dimensions that could be explored for future research.

Related Work

Spamming is not only prevalent in OSN's but is widespread in other online domains as well and one such domain is online review markets where spamming is done by leaving unethical reviews. Hence, in [3] the author proposed a feature framework to address the fake review identification problem in the consumer electronic good domain. A dataset was constructed and the appropriate feature set was identified and proposed. Using the proposed F3 framework, the method could attain 83% accuracy in isolating the fake reviews. In [4], the author experimented with the distribution of observations in the training and testing dataset of twitter spam profiles and concluded that unequal distribution of training and test set has a great impact on classification accuracy rate. A fuzzy oversampling method was employed to generate synthetic data samples from the observed samples. In [5], the author proposed a method based on unsupervised clustering that attempted to analyze the users' reactions through smileys. The reactions were saved and by applying similarity measures and unsupervised clustering techniques, they further classified the spammers. As the reactions are instantaneous, an investigation of these reactions delivered vital information to find abnormal activities on Facebook accounts.

Many spam detection methods have been suggested in the early works for identifying the spam profiles as well as spam messages. The application of these techniques was successful only to some extent because the spammers frequently change their strategies of evasion techniques that led to the deterioration of the identifier systems. In [6], the author proposed a dynamic metric to quantify the transformation in user actions by considering time-based factors and designed a new set of features to compute user evolution patterns.

The classifier demonstrated a better performance upon considering the features that define the user evolution pattern. In [7], the author considered the Sina Weibo platform for identifying spammers. Apart from the user and social related information, the entire message content was incorporated for the problem. In [8], the author considered the problem of spam detection in tweets using language features, a character n-gram in spam detection that differs from the existing detection techniques on detecting the user account. The computational performance of the proposed method revealed that 1K tweets could be effectively classified within a second by using the features proposed.

In [9], the author proposed an optimized set of features independent of historical tweets, which are only available for a short period on Twitter to classify twitter profiles as spam and non-spam. Features related to the users, accounts, and pair-wise engagement with each other are employed. The effectiveness of the method was proven by comparing it to a classical feature set for spam detection in the literature. In [10], the author proposed a unique structure for spam detection in Twitter that employed probabilistic data structures as machine learning algorithms in various stages of an ensemble to recognize a tweet as a Spam or Ham. A majority voting scheme was incorporated to arrive at the final decision of the four-stage ensemble. In [11], the author designed a hybrid machine learning technique based on Support Vector Machines and Whale Optimization Algorithm for the task of categorizing spammers in online social networks. The proposed model executed the automatic discovery of spammers and provided an understanding of the most influential features that influence the problem. Moreover, the efficacy of the model was verified on four lingual datasets in four different languages namely Arabic, English, Spanish, and Korean, collected from Twitter.

In [12], the authors designed a unified spam identification framework to locate spammers on Twitter and Facebook utilizing the most common mutual features of both the OSN's in terms of user posts, conduct and user participation, etc. In [13], a unified methodology called SSCF was proposed, to sense spammers in Sina Weibo, a popular social network in China. The approach considers the multi-view features and recommends a solution that

can merge the comprehensive clues discovered from these views to recognize spammers. Though several methods were proposed in the literature to overcome the challenges posed by spammers, it stands perplexing by continuously posing challenges from various dimensions. The population-based methods have been proved to be efficient searching techniques. Hence, WOA being a global optimizer and a proven effective searching strategy is considered in this work and an improved version of the same is presented later in the paper.

The WOA has been improved and modified for various other problems in the literature as in [19], [20], [21], [22], [23], [24] and [26].

Basics and Background

A. The Whale Optimization Algorithm

The Whale optimization algorithm (WOA) was essentially developed influenced by the foraging behavior of humpback whales [15]. These whales exhibit a mechanism called "Bubble net feeding" to trap the prey in a circle of bubbles. The working of this algorithm involves two phases. (1) Exploration – accomplished by randomly searching for the prey (2) Exploitation – achieved through actions such as encircling the prey and spiral bubble net attacking mechanisms. The encircling mechanism is mathematically formulated using eq. (2) to change the position of whales.

$$\vec{D} = \vec{C}.\overrightarrow{X^*}(t) - \vec{X}(t) \quad (1)$$

$$\vec{X}(t+1) = \overrightarrow{X^*}(t) - \vec{A}.\vec{D} \quad (2)$$

where 't' denotes the present iteration, X^* denotes the best solution found so far, X is the current solution, A & C are co-efficient vectors and are calculated using the equations (3) & (4),

$$\vec{A} = 2\,\vec{a}.\vec{r} - \vec{a} \quad (3)$$

$$\vec{C} = 2.\vec{r} \quad (4)$$

$$a = 2 - t\frac{2}{t_{Max}} \quad (5)$$

where a reduces from 2 to 0 for iterations and r is a random vector in the range [0,1], t is the current iteration, t_{Max} is the maximum iteration preset for the procedure. As can be

seen in the process of encircling prey, the positions of the whales are modified based on the position of the best food source or the whale that possesses the best fitness value. The whales traverse towards the krill (prey) in a dwindling encircling pattern as well as a spiral-shaped track. The movement of the spiral-shaped track is formulated as follows,

$$\vec{X}(t+1) = \overrightarrow{D'}.e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t) \quad (6)$$

$$\overrightarrow{D'} = |\overrightarrow{X^*}(t) - \vec{X}(t)| \quad (7)$$

where $\overrightarrow{D'}$ is the distance between the i^{th} whale and target, b is a constant that defines the shape of the spiral & l is a random number in the range [-1, 1]. The humpback whales travel around the prey as well as migrate along a spiral track simultaneously. This concurrent behavior is replicated by introducing a probability that assumes 50% to choose between the encircling mechanism and the spiral track updating procedure to change the positions.

$$\vec{X}(t+1) = \begin{cases} \overrightarrow{X^*}(t).\vec{A}.\vec{D} & p < 0.5 \\ \overrightarrow{D'}.e^{bl}.\cos(2\pi l) + \overrightarrow{X^*}(t) & p \geq 0.5 \end{cases} \quad (8)$$

where p is a random number in [0, 1]

Besides exhibiting the aforesaid mechanisms for updating the positions, the whales also perform a random search to discover the prey that constitutes the exploration phase or global search of this algorithm [16]. A whale is randomly chosen and the positions of the rest of the whales are adjusted based on this randomly chosen whale. The condition $|\vec{A}|>1$ greatly facilitates the global search process and is executed using the equation (10)

$$\vec{D} = |\vec{C}.\overrightarrow{X_{rand}} - \vec{X}| \quad (9)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A}.\vec{D} \quad (10)$$

where $\overrightarrow{X_{rand}}$, is a whale randomly chosen from the population.

As Jun Luo [19] pointed out, though WOA is a global optimizer possessing the capability of exploration and exploitation, the algorithm loses the exploration ability for higher iterations and stagnate with subsequent periodical exploitative generations. The algorithm lacks mechanisms to escape from local optima. Slow convergence and entrapping in local optima are the limitations of WOA.

B. The Salp Swarm Algorithm

The Salp Swarm algorithm was proposed by *Seyedali Mirjalili* [20] inspired by the foraging behavior of salps in the sea. The exploitation ability and simplicity of SSA has motivated us to embed it in the WOA. The salps are transparent, barrel-shaped, and resemble jellyfishes in appearance. The locomotion of salps is accomplished by contracting and squirting water through their transparent body. The swarm exhibits a unique pattern during the foraging process by flocking themselves into a chain structure. Furthermore, the swarm of salps consists of two distinct groups' viz. (i) leader and (ii) followers. The foremost salp in the salp chain is designated as a leader while the rest of them succeeding are selected as followers.

The salps move in search of the best food source in *n* dimensions, where 'n' represents the problem variables. The positions of these salps are to be updated periodically until the best food source is found. This behavior is modeled in a mathematical form to solve real-world problems. The position of the leader is updated using equation (11).

$$x_j^1 = \begin{cases} F_j + C_1\left((ub_j - lb_j) \times C_2 + lb_j\right) & C_3 \leq 0 \\ F_j - C_1\left((ub_j - lb_j) \times C_2 + lb_j\right) & C_3 > 0 \end{cases} \qquad (11)$$

where x_j^1 represents the first salp, the leader, F_j represents the position of the best food source, ub_j and lb_j denotes the upper and lower bounds respectively. C_2 and C_3 are random variables in the range [0, 1]. C_1 is calculated using the following equation,

$$C_1 = 2e^{-\left(\frac{4t}{t_{Max}}\right)^2} \qquad (12)$$

where *t* and t_{Max} denotes the present and maximum iterations set in the algorithm. The position of the followers is updated using equation (13) and is as follows,

$$x_j^i = \frac{1}{2}\left(x_j^i + x_j^{i-1}\right) \quad (1 < i < m), (1 \leq j \leq n) \qquad (13)$$

where x_j^i represents the i^{th} follower of the chain in the j^{th} dimension, *m* indicates the total number of salps in the population, and *n* represents the dimensions considered in the problem.

The Proposed Hybrid Modified Whale Optimization Algorithm for Spam Profile Detection (MWOA-SPD)

The feature selection process helps in identifying the most significant features and also eliminates redundant or noisy ones with the sole objective of reducing dimensions and enhancing prediction accuracy [18].

A. Solution Representation

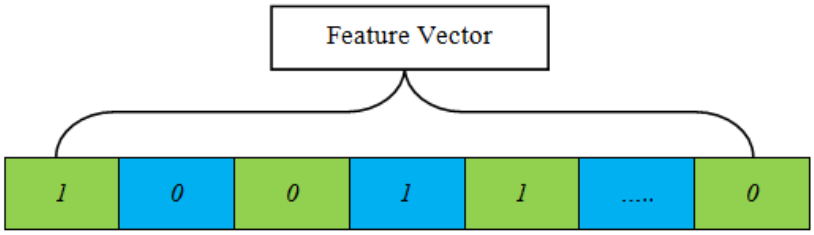

Fig.1 Representation of a whale

Initial Population =	$X\backslash F$	F_1	F_2	F_3	F_4	...	F_n
	X_1	1	0	1	1	...	0
	X_2	0	0	1	0	...	1
	:	:	:	:	:	:	:
	X_m	1	1	0	0	...	1

Fig.2 Illustration of a population

The presence of relevant features affects the classification accuracy and hence, necessary information would be concealed if these contributing features are removed. Various nature-inspired algorithms are being employed to address global optimization problems. And certainly, a challenging task is to determine the optimal set of features without losing the classification accuracy, particularly for larger datasets.

In general, the spam profile classification problem consists of several features and not all of these will contribute to the process. It is to be noted that some features, if not eliminated, deteriorates the performance of the entire system. Similarly, the presence of certain features greatly contributes to the classification process. The proposed hybrid M-WOA exhibits better performance than the traditional WOA in terms of better results and faster convergence rate.

M-WOA generates solutions with a vector of binary values. As illustrated in Fig. 1 & Fig. 2, the presence of a '1' indicates the selection of a feature, and a '0' indicates that the feature is not included in the set. The length of the vector equals the number of features of the problem in hand and the number of vectors corresponds to the initial population in the algorithm. For each whale, a solution vector is randomly generated with binary values [0, 1]

B. Fitness Evaluation

After the solutions are generated, each feature set is evaluated to test its effectiveness. In our work, the accuracy of the classifier is used as a fitness value to evaluate these individual solutions.

$$\text{Accuracy}=\frac{TP+TN}{TP+FN+FP+TN} \qquad (14)$$

TABLE I. Metrics to find accuracy

Metric	Description
True Positive (***TP***)	No. of Twitter profiles that have been correctly categorized as a spam class
False Positive (***FP***)	No. of Twitter profiles that have been miscategorized as a spam class
True Negative (***TN***)	No. of Twitter profiles that have been correctly categorized as a non-spam class
False Negative (***FN***)	No. of Twitter profiles that have been miscategorized as a non-spam class

Accuracy is calculated using equation (14) and is directed back to MWOA for further processing. The values TP, TN, FP, *and* FN *are determined using the calculations provided in Table 1.*

C. Scheme Architecture

The MWOA- hybrid technique involves a series of steps to find the best arrangement of features without compromising the accuracy of the classifier. After completing the pre-processing steps, the dataset is to be loaded. As illustrated in Fig. 3, the MWOA initiates the process by generating random solutions for each whale with either a '1' or '0'. The dataset is then equally split into training and a testing set. The features that are marked as '1' are considered for the training and testing purpose. The classifier is then trained using the training data and tested for accuracy. The process is repeated until the fitness value is

calculated for all the search agents. The parameters *A, a, C, D* are initialized and a random value p is generated in the range [0, 1].

Based on this probability, the algorithm chooses to switch between an encircling mechanism and a spiral updating mechanism of WOA to update the position of solutions. If p<0.5, the method can either update the position using the encircling movement of whales or the methodology followed by the salps to update the position depending on the value of |A|. The population used by WOA is shared by salps and appropriately the positions are updated. The leader of the salp chain migrates to the best food source available whereas, the followers in the chain change their position based on the foregoing salps. If the termination condition is not met, the phases of feature subset selection, fitness evaluation, and repositioning of the search agents will be iterated. If the termination criterion holds, the algorithm outputs the best solution obtained among all iterations and runs, in the form of a vector that corresponds to the selected features.

The SSA algorithm offers the benefit of both diversification and intensification. The randomization in the SSA guides the process to search for the best food source in diverse dimensions. As the leader always moves towards the best solution, the exploiting ability of the optimization algorithm is also achieved. The SSA has been applied in several optimization problems due to its simplicity and efficiency [26], [27], [28], [29] & [30].

Algorithm 1 The Pseudocode of Modified WOA Algorithm for Spam Profile Identification	
1	Generate Initial population of whales W*i* (*i* = 1, 2,.., m)
2	**for** each solution **do**
3	Evaluate fitness
4	X^* = the best fit whale
5	**while** *t* < *Max* _Iterations **do**
6	**for** each solution **do**
7	Update *a*, D, *A*, *C*, *l*, *and p*
8	**if** *p* < 0. 5 **then**
9	**if** | *A* | < 1 **then**
10	Update the position of the current whale using Eq. (2)
11	**else if** | *A* | ≥ 1 **then**
12	if X_i == leader
13	Modify the position of the current whale using Eq. (11)
14	**else if** X_i ==follower **then**
15	Modify the position of the current whale using Eq. (13)
16	**else if** *p* ≥ 0. 5 **then**
17	Update the position of the current whale using Eq. (6)
18	Calculate the fitness of each whale
19	Adjust X* if there exists a better solution
20	t = t + 1
21	**return** X*

The limitation of WOA is overcome using this hybrid technique. The exploration process of this method diversifies the population thereby producing a large number of varied solutions. Replacing the exploration methodology of WOA with the position updating mechanism of SSA further intensifies the exploration and exploitation thus reducing the iterations to attain global optimum. The leader of the salp chain always moves towards the best food source thus alleviating the problem of trapping in the local optima. As the followers of the salp chain follow the leader, paves a way to escape from the local optimal position and also guarantees to yield the best possible solution. An increase in the randomization may degrade the accuracy of the solution [17]. To get rid of this condition, the exploitation happened only around the best solutions obtained so far.

Experimental Procedure

A. Dataset Preparation

To perform the spam profile classification problem, a twitter dataset was manually constructed, and using the oversampling techniques, a balanced dataset was created with 9,688 instances. The details of the dataset are provided in Table II.

TABLE II. Dataset Details

Dataset Description	
No. of Instances	9688
No. of Attributes	26
No. of Classes	2

B. Feature Framework

Attributes related to user-profile, user-activity, and account are employed for the classification procedure. Furthermore, we propose a novel set of features to enhance spam detection.

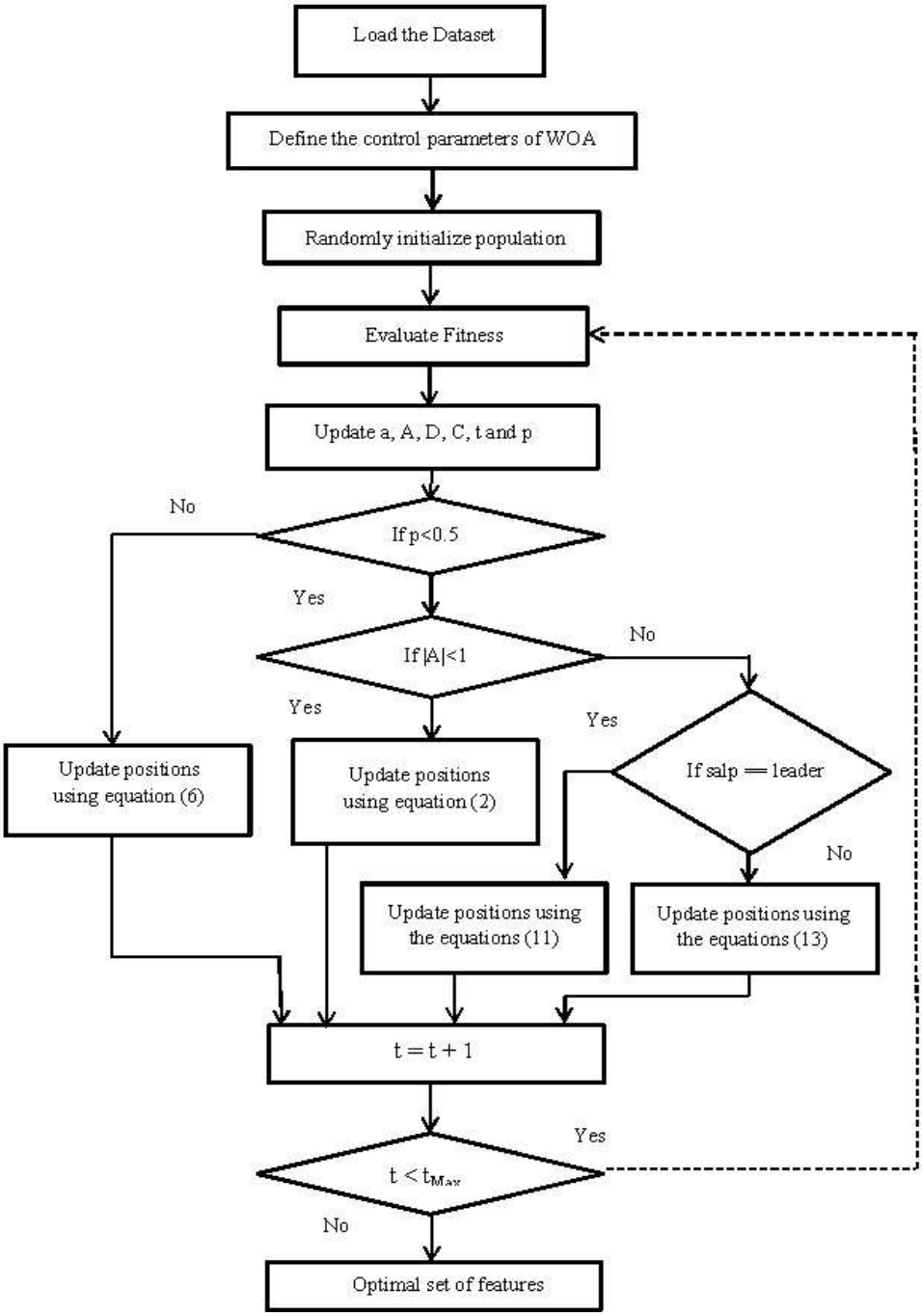

Fig.3 Flow chart of MWOA-SPD

Besides having several algorithms for addressing the spammers, the reason that all of them were unsuccessful is due to the evolving nature of spammers. A method devised during a particular period may not be effective during another course of time as the spammers keep evolving and altering the strategies to evade the network of trust. Hence, there is a pressing need to have detector systems that would be robust with time and

locates the spammers in the heavily occupied social network. The IP addresses are numerical labels that are assigned to devices connected to the network [31] & [32]. The values of IP addresses cannot be modified, manipulated, or camouflaged unlike features such as no. of followers, tweets, URLs, etc., Thus, analyzing the IP address based activities would simplify the task of uncovering the adversaries hidden in the network. The spammers exhibit a unique pattern that could visibly be traced using the IP address based features. Table IV shows the proposed feature set F_{22}, F_{23}, F_{24}, F_{25}, and F_{26}. Based on an extensive study and analysis from the literature, a complete feature set considered for the problem is listed in Table III.

TABLE III. List of features used

List of Features	
F_1	Twitter Account No.
F_2	Account age
F_3	No. of followers
F_4	No. of followings
F_5	No. of user favorites
F_6	No. of lists
F_7	No. of tweets
F_8	No. of re-tweets
F_9	No. of hashtags
F_{10}	No of user mentions
F_{11}	No of URLs
F_{12}	Repeated Words
F_{13}	Words in capital
F_{14}	The frequency of function words
F_{15}	Special character
F_{16}	Emoticons
F_{17}	Digits
F_{18}	Minimum time between Tweet postings
F_{19}	Maximum time between Tweet postings
F_{20}	Tweets Posted Per Day
F_{21}	Tweets Posted Per Week
F_{22}	IP address changed last 1 week
F_{23}	IP address changed 24 hours
F_{24}	No. of IP address for posts
F_{25}	No. of IP address for comments
F_{26}	No of Tweets deleted

TABLE IV. Description of the Proposed Features

Feature	Description
F_{22}	Total no. of unique IP addresses recorded from the login activity during the last 1 week
F_{23}	Total no. of unique IP addresses recorded from the login activity during the last 24 hours
F_{24}	No. of unique IP addresses from Tweet postings
F_{25}	No. of unique IP addresses from comment activity
F_{26}	No. of Tweets deleted

C. Evaluation Measures

The evaluation metrics employed in this work to qualify the objectives include *Precision, Recall*, *F-Measure*, and *Accuracy.* As mentioned in Section 4, the accuracy is used as the fitness function and the remaining three metrics are determined as follows:

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

$$F\text{-}Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (17)$$

D. Experiment Setup

All the algorithms and classifiers employed in this work have manually been coded in Python language using Spyder text editor, a free integrated development environment included with Anaconda. The experiments were performed on a Windows 10 machine with Intel Core i7-3630, 2.40 GHz processor and 8 GB RAM.

The experiment is performed in three stages. Three structures of neural networks namely Artificial Neural Network (ANN), Deep Learning Neural Network (DNN), and Recurrent Neural Network (RNN) are used for class labeling. In the first phase, the Neural Network based classifiers DNN, ANN, and RNN are applied to the dataset without performing feature selection. The neural networks have showcased fair performance for real-world datasets [33]. In the second phase, the classical WOA in conjunction with these three variants of neural network classifiers is used to analyze the outcome. And finally, the proposed MWOA with the NN variants are evaluated on the dataset to compare the performance. The parameter settings of the techniques considered are provided in Table V. The forward propagation neural network has the number of input nodes set to the number of variables. The Relu activation function is used in the hidden layer and Sigmoid on the output layer. Three hidden layers with 26, 12, and 2 neurons respectively, were

employed with a learning rate of 0.05% and a batch size of 32. The ratio of the training and testing data is 70:30.

TABLE V. Description of the Proposed Features

Parameter	**Value**
Whales	6
No. of Iterations	15,20,25
No. of Runs	5
Boundary values	Binary vector [0, 1]
Salps	Same as whales

Evaluation Results

TABLE VI. Description of the Proposed Features

	Accuracy		
Method	Epochs = 100	Epochs = 200	Epochs= 300
ANN	50.49	50.64	51.33
DNN	50	48.54	50.11
RNN	63.17	62.11	**63.59**

In the initial phase, we tested the dataset with ANN, DNN, and RNN without incorporating any feature selection algorithm. The results are presented in Table VI. It could be seen that the accuracy is only about 50% when classified without selecting the significant features. RNN performs considerably better than the other classifiers when considering all the features for the classification process.

TABLE VII. Evaluation results when whales = 6

Method	Iteration	Accuracy	Precision	Recall	F- Measure	Features Selected
WOA + ANN	15	71.77	0.7266	0.7053	0.7158	13
	20	68.07	0.7288	0.7235	0.7261	15
	25	72.92	0.7123	0.7215	0.7169	13
WOA + DNN	15	70.89	0.7174	0.7063	0.7118	16
	20	75.74	0.7631	0.7644	0.7637	20
	25	78.07	0.7863	0.7922	0.7892	18
WOA + RNN	15	56.34	0.5600	0.5785	0.5691	12
	20	59.29	0.6047	0.5991	0.6019	11
	25	52.85	0.5386	0.5568	0.5475	13
MWOA-SPD + ANN	15	84.35	0.8422	0.8569	0.8495	11
	20	85.14	0.8522	0.8316	0.8418	13
	25	**85.23**	**0.8874**	**0.8896**	**0.8885**	13
MWOA-SPD + DNN	15	83.35	0.8025	0.8709	0.7064	17
	20	80.98	0.7439	0.7252	0.7245	13
	25	77.67	0.7515	0.7397	0.7406	11
MWOA-SPD + RNN	15	54.23	0.5009	0.5131	0.5069	15
	20	54.00	0.5382	0.5259	0.5320	16
	25	57.46	0.5524	0.5667	0.5595	12

The Table VII displays the results of the three network structures with feature selection performed by the conventional WOA and proposed MWOA-SPD. The number of whales has been set to 6 and evaluated for different iterations. For 15 iterations, the MWOA-SPD + ANN gives the best results with an accuracy of 84.35% followed by MWOA-SPD + DNN and WOA + ANN with 83.35% and 71.77% respectively. For 20 and 25 iterations, the MWOA-SPD + ANN achieves the best result with a minimum feature set of 13 attributes. Though this combination of algorithm yields a lowest number of 11 features for 15 iterations, considering the variation in the accuracy, this is discarded. For 20 iterations, the MWOA-SPD + DNN and WOA + DNN stand second and third in discriminating the spam profiles with an accuracy of 80.98% and 75.74% respectively. However, for 25 iterations, the second and third positions of these algorithms are swapped with 78.07% and 77.67% respectively. The MWOA-SPD + ANN yields best results in terms of accuracy, precision, recall and F-Measure when the number of whales is set to 6. Though many combinations of techniques in the above table selected minimal subset, the accuracy should be given due importance in association with the minimal subset. The table VIII displays the outcomes when the number of initial populations is set to 8. Similar to the previous case, the MWOA-SPD + ANN yields the best result with an accuracy of 86.78%. Though it has obtained a best precision value of 0.8733 during 20 iterations, it is still lesser than value value obtained when the number of whales was set to 6. The Recall and F-Measure are higher in this case where the number of whales is set to 8. The algorithm yields the same feature count as that of the previous count as 13. The MWOA-SPD + DNN yields competitive results followed by WOA + DNN with 73.95%. Similarly the second best precision has been obtained MWOA-SPD + DNN followed by WOA + DNN with 0.8152 and 0.7962 respectively.

Likewise, these two techniques occupy second and third positions for recall with 0.8035 and 0.7997 respectively. It can be concluded that the MWOA-SPD + ANN outperforms the other techniques for 20 iterations when the number of whales is set to 8. The RNN doesn't show improvements upon performing feature selection using either WOA or MWOA-SPD. However, there is a notable increase in the performance when feature selection is performed on the Twitter dataset. The ideal feature set output by the algorithm is 13.

TABLE VIII. Evaluation results when whales = 8

Method	Iteration	Accuracy	Precision	Recall	F-Measure	Features Selected
WOA + ANN	15	72.06	0.7301	0.7104	0.7201	14
	20	68.66	0.7221	0.7341	0.7281	16
	25	71.23	0.7456	0.7117	0.7283	11
WOA + DNN	15	71.44	0.7226	0.7002	0.7112	17
	20	73.95	0.7591	0.7322	0.7454	15
	25	72.58	0.7962	0.7997	0.7979	18
WOA + RNN	15	55.23	0.5646	0.5324	0.5480	13
	20	57.87	0.6282	0.6173	0.6227	12
	25	54.44	0.6154	0.6152	0.6153	12
MWOA-SPD + ANN	15	83.56	0.8066	0.8255	0.8159	11
	20	**86.78**	0.8725	**0.9003**	**0.8862**	13
	25	83.89	**0.8733**	0.8544	0.8637	12
MWOA-SPD + DNN	15	82.29	0.8152	0.8035	0.8093	15
	20	81.36	0.7526	0.7865	0.7692	13
	25	76.82	0.7797	0.7524	0.7658	14
MWOA-SPD + RNN	15	52.54	0.5496	0.6055	0.5762	15
	20	60.21	0.5877	0.5583	0.5726	12
	25	54.63	0.5634	0.5426	0.5528	16

Conclusion and Future Work

This paper presented a hybrid wrapper whale optimization algorithm combining WAO with another evolutionary approach, SSA. The utilization of SSA in the proposed technique prevents the algorithm from trapping into local optima and facilitates in reaching global optimum. The technique was tested on a Twitter dataset that was manually constructed. The experimental results demonstrate the effectiveness and robustness of the MWOA-SPD technique to the Spam Profile Detection problem. The work presented here attempts to identify the spammers on Twitter. As the spammers constantly change their strategies of spamming, we further proposed a robust set of features that would be difficult to manipulate or escape from.

Experimental results also reveal that the proposed technique is efficient in selecting a relevant smaller subset of features with high classification accuracy. Further, the performance of the proposed method is also compared to traditional WOA, standalone neural network classifiers, and combinations of them.

It is evident from the results that the integration of WOA with SSA significantly improved the performance of the classifier. Due to intensive exploration and exploitation, the

approach was very effective in handling this high dimensional problem. However, further research would be carried out in the future to fine-tune the system to be effective for the social networks of all kinds.

References

https://www.oberlo.in/blog/social-media-marketing-statistics

Soman SJ. A survey on behaviors exhibited by spammers in popular social media networks. In International Conference on Circuit, Power and Computing Technologies (ICCPCT) 2016: 1-6.

Barbado R, Araque O, Iglesias CA. A framework for fake review detection in online consumer electronics retailers. Information Processing & Management 2019; 56(4): 1234-1244.

Liu S, Wang Y, Zhang J, Chen C, Xiang Y. Addressing the class imbalance problem in twitter spam detection using ensemble learning. Computers & Security 2016; 69: 35-49.

Savyan PV, Bhanu SMS. Behaviour profiling of reactions in facebook posts for anomaly detection. In Ninth International Conference on Advanced Computing (ICoAC) 2017: 220-226.

Fu Q, Feng B, Guo D, Li Q. Combating the evolving spammers in online social networks. Computers & Security 2017; 72: 60-73.

Yu D, Chen N, Jiang F, Fu B, Qin A. Constrained NMF-based semi-supervised learning for social media spammer detection. Knowledge-Based Systems, 2017; 125: 64-73.

Ashour M, Salama C, El-Kharashi MW. Detecting Spam Tweets using Character N-gram Features. In 13th International conference on computer engineering and systems (ICCES) 2018: 190-195.

Inuwa-Dutse I, Liptrott M, Korkontzelos I. Detection of spam-posting accounts on Twitter. Neurocomputing 2018; 315: 496-511.

Singh A, Batra S. Ensemble based spam detection in social IoT using probabilistic data structures. Future Generation Computer Systems 2017; 81: 359-371.

Ala'M AZ, Faris H, Alqatawna JF, Hassonah MA. Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts. Knowledge-Based Systems 2018; 153: 91-104.

Pandey J, Job MA. Proposed framework for Spam recognition in big data for Social Media Networks in smart environment. In 4th MEC International Conference on Big Data and Smart City (ICBDSC) 2019: 1-5.

Chen H, Liu J, Lv Y, Li MH, Liu M, Zheng Q. Semi-supervised clue fusion for spammer detection in Sina Weibo. Information Fusion, 2017; 44: 22-32.

Soliman GM, Abou-El-Enien TH, Emary E, Khorshid MM. A Hybrid Modified Whale Optimization Algorithm with Simulated Annealing for Terrorism Prediction. Ingénierie des Systèmes d Inf., 2019; 24(3): 281-287.

Mirjalili S, Lewis A. The whale optimization algorithm. Advances in engineering software 2016; 95: 51-67.

Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Applied Soft Computing 2018; 62: 441-453.

Abdel-Basset M, Manogaran G, El-Shahat D, Mirjalili S. A hybrid whale optimization algorithm based on local search strategy for the permutation flow shop scheduling problem. Future Generation Computer Systems 2018; 85; 129-145.

Sayed GI, Darwish A, Hassanien AE. A new chaotic whale optimization algorithm for features selection. Journal of classification, 2018; 35(2): 300-344.

Luo J, Shi B. A hybrid whale optimization algorithm based on modified differential evolution for global optimization problems. Applied Intelligence, 2018; 49(5): 1982-2000.

Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. Advances in Engineering Software, 2017; 114: 163-191.

Zheng Y, Li Y, Wang G, Chen Y, Xu Q, Fan J, Cui X. A novel hybrid algorithm for feature selection based on whale optimization algorithm. IEEE Access 2018; 7: 14908-14923.

Kaur G, Arora S. Chaotic whale optimization algorithm. Journal of Computational Design and Engineering, 2018; 5(3): 275-284.

Mostafa Bozorgi S, Yazdani S. IWOA: An improved whale optimization algorithm for optimization problems. Journal of Computational Design and Engineering 2019; 6(3): 243-259.

Jiang T, Zhang C, Zhu H, Gu J, Deng G. Energy-efficient scheduling for a job shop using an improved whale optimization algorithm. Mathematics 2018; 6(11).

Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection. Applied Soft Computing 2018; 62: 441-453.

Hegazy AE, Makhlouf MA, El-Tawel GS. Improved salp swarm algorithm for feature selection. Journal of King Saud University-Computer and Information Sciences 2020; 32(3): 335-344.

Aljarah I, Mafarja M, Heidari AA, Faris H, Zhang Y, Mirjalili S. Asynchronous accelerating multi-leader salp chains for feature selection. Applied Soft Computing 2018; 71: 964-979.

Ibrahim HT, Mazher WJ, Ucan ON, Bayat O. Feature selection using salp swarm algorithm for real biomedical datasets. IJCSNS, 2017; 17(12): 13-20.

Ibrahim RA, Ewees AA, Oliva D, Abd Elaziz M, Lu S. Improved salp swarm algorithm based on particle swarm optimization for feature selection. Journal of Ambient Intelligence and Humanized Computing 2018; 10(8): 3155-3169.

Sobhi A, Majdi M, Hossam F, Ibrahim A. Feature Selection Using Salp Swarm Algorithm with Chaos, ISMSI ’18, 2018, Phuket, Thailand.

Ala'M AZ, Alqatawna JF, Faris H. Spam profile detection in social networks based on public features. In 8th International Conference on information and Communication Systems (ICICS) 2017; 130-135.

Inuwa-Dutse I, Liptrott M, Korkontzelos I. Detection of spam-posting accounts on Twitter. Neurocomputing 2018; 315: 496-511.

Arram A, Mousa H, Zainal A. Spam detection using hybrid Artificial Neural Network and Genetic algorithm. In 13th International Conference on Intellient Systems Design and Applications 2013: 336-340.

Kiani M, Asemi A, CheshmehSohrabi M, Shabani A. Information ecology of bioinformatic in web of science with emphasizing on articles thematic interaction. Webology 2020; 17(1) (A215): 171-190.

Shenavar A, Doulani A. Review of iranian journal articles indexed in web of science based on altmetric indicators in scientific social media. Webology 2020; 17(1) (A214): 158-170.

Author Biography

Mrs. R. Krithiga currently works as an Assistant Professor, Department of Computer Applications at Perunthalaivar Kamarajar Arts College, Madagadipet, Puducherry. She completed her Master of Computer Science from Pondicherry University. She is also a research scholar at Pondicherry Engineering College. Her area of interest includes Evolutionary algorithms, Data mining, and Intelligent systems.

Dr.E. Ilavarasan is a Professor of the Department of Computer Science & Engineering at Pondicherry Engineering College, Puducherry. He has more than 25 years of experience in the teaching field. He is an expert in Web service computing.