Wireless Pers Commun (2018) 103:1353–1374
https://doi.org/10.1007/s11277-018-5513-z

Detecting Streaming of Twitter Spam Using Hybrid Method

N. Senthil Murugan[1] · G. Usha Devi[1]

Published online: 27 February 2018

Abstract Twitter, the social network which evolving faster and regular usage by millions of people and who become addicted to it. So spam playing a major role for Twitter users to distract them and grab their attention over them. Spammers actually detailed like who send unwanted and irrelevant messages or websites and promote them to several users. To overcome the problem many researchers proposed some ideas using some machine learning algorithms to detect the spammers. In this research work, a new hybrid approach is proposed to detect the streaming of Twitter spam in a real-time using the combination of a Decision tree, Particle Swarm Optimization and Genetic algorithm. Twitter has given access to the researchers to get tweets from its Twitter-API for real-time streaming of tweet data which they can get direct access to public tweets. Here 600 million tweets are created by using URL based security tool and further some features are extracted for representation of tweets in real-time detection of spam. In addition, our research results are compared with other hybrid algorithms which a better detection rate is given by our proposed work.

Keywords Decision tree · Particle swarm optimization · Genetic algorithm · Feature extraction and machine learning

1 Introduction

In this new era the development of social networks like Twitter, Facebook etc. have become ultimate usage for humans in their day-to-day life. As we can see in today's modern world more than 0.31 billion people are using Twitter, and still increasing many users. So people expecting to share or communicate their messages through this Twitter in

✉ N. Senthil Murugan
senthilmurugan.n@vit.ac.in

G. Usha Devi
ushadevi.g@vit.ac.in

[1] School of Information Technology and Engineering, VIT University, Vellore, India

a safe and secure manner. In this generation many leaders of several countries and actors, also several business people are sharing their ideas and messages through this microblogging site. Twitter has become the most popular social blog for many people and more than 600 million tweets are tweeting per day. This reflects many Twitter users which lead to spammers to interfere in terms of unwanted messages, videos, images etc [1, 2]. Many users become victim to these spammers not even aware of the spam. In 2013, an email is sent much like phishing of messages to the election commission of Australia which conciliate to the spam messages sent to their account.

For this reason, researchers proposed many algorithms for the detection of spam and twitter itself set some rules to defend the spammers like deactivation of accounts who misuses, sending identical messages. Twitter set some options to their users to report about any spam account which been acting abnormal by way of their behavior. Several researchers have developed many tools to detect the spam automatically by using some machine learning algorithms by classifying the problem as per their expectation. One such example is the system known as Web Reputation Technology used by Security Company named Trend Micro who boycotts service for its users to filter the spam URLs [3]. A blacklist filter is developed by twitter for detecting spam and named as botmaker [4]. Because of time failure, botmaker could not able detect the recent spams which fail to safeguard the twitter users [5]. Unfortunately more than 90% victims gone through some new spam before it could be detected by the botmaker [6]. The researchers have proposed some ML algorithms to find particulars for the restrictions of blacklist by using statistical features of 'spam tweets' without examining the URLs [7, 8].

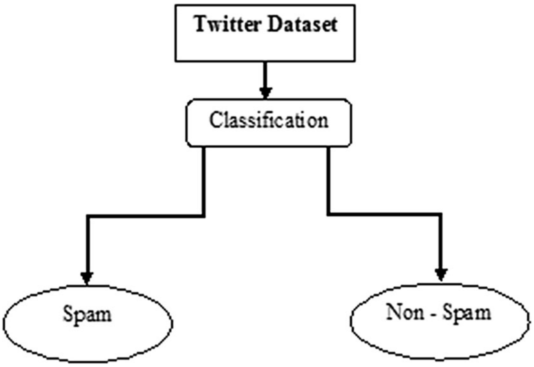
Machine learning algorithms played a major role in finding twitter spam; researchers almost use the concept of ML algorithms for the detection of spam but still spammers are getting new way to avoid detecting. So for this reason many researchers proposing new solution to overcome the problem faced by the spammers. Twitter facing regular problem when streaming of tweets in a day hacked by spammers and unable to detect much easily and even accuracy of detection using ML algorithms unsuccessful to overwhelm the issues faced by the twitter users. In this paper we proposed a combination of Decision tree with the Genetic algorithm to increase the detection rate of spam [9–11]. We collected 6.5 million tweets using twitters streaming API and categorizing the spam using our proposed algorithm, then to improve the detection rate using Evolutionary algorithms commonly used Particle Swarm Optimization (PSO), Decision Tree (DT) and Genetic Algorithm (GA). At last a comparison of the proposed work with some of the main hybrid algorithms is shown and our research results given promising outcome when using the combination of three methods as PSG-DT.

1.1 Decision Tree

It is the decision making implementation by using the general behavior of tree structure. Decision tree analyze possible consequences of utility by the given resource. Many researchers have studied how to analyze the spam tweets and classify it in which one of its machine learning algorithms named as decision tree. Figure 1 shows the simple structure of decision tree.

Decision tree is a family of supervised learning algorithms which played a basement role for every researcher to compete for the detection of spam in twitter. The accuracy level for the detection rate is high but consecutively having problem while steaming of spam flow continuously while detecting. So far the researchers concluded that decision tree would give a possible result for classifying spam and non-spam in twitter. Unfortunately

Fig. 1 Decision tree example



the improvement for the detection rate is still a challenging task for researchers to find a good solution [12, 13].

Types of Decision Trees:

1. Absolute variable

The predefine variables given to solve the problem or to predict the solution by having past information.

2. Continual variable

Here the variable is continuous which has to predict the target by analyzing other variables given by the source.

1.2 Genetic Algorithm

It is defined as a higher level of procedure to learn something by them. One of the most common used optimization techniques and it is the part of evolutionary algorithms. Mutation, Crossover, and Selection are three main concepts used to obtain high quality solutions to optimization and search problems. The Fig. 2 shows the architecture of GA.

The GA process takes a deep integration of development over the given data. Starting it utilizes a solution to the given group of data which is known as population and initializes the data into random or heuristic method to get a conscious decision or learn a new solution by them. Next it finds a fitness evaluation for the given population to get the optimal solution. If predicted output is not upon the expectation then crossover process is taken place where it performs a comparison of the processed data and mutation begins convergence to get the proper solution while crossover fail to perform well. After completing the steps, reproduction of obtained data could be analyzed to regenerate from starting if the best solution was not obtained and also termination of process taken when the outcome of prescribed or close solution is optimal [14, 15].

1.3 Particle Swarm Optimization

Dr.Eberhart and Dr.Kennedy developed a new population based optimization technique which motivated by the concept of fish schooling or bird flocking by using their social behavior. It is a homogenous with existing evolutionary techniques like genetic algorithm.

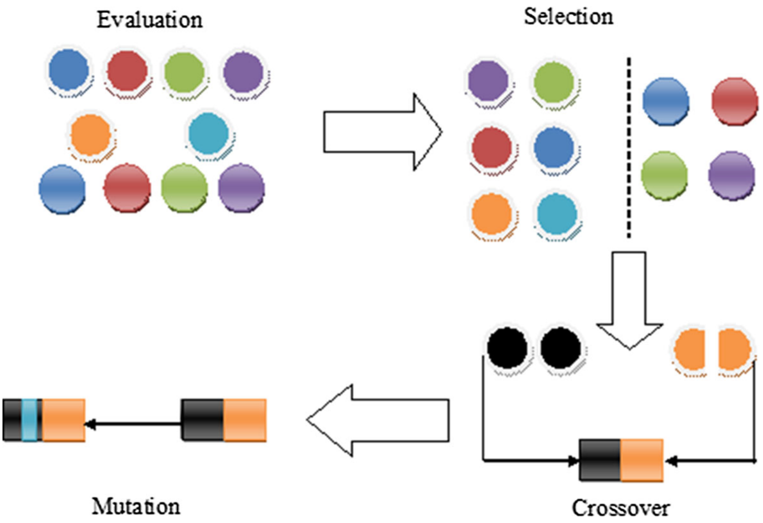


Fig. 2 Architecture of genetic algorithm

Unlike GA the PSO has no development operators such as mutation and crossover; it actually works with the movement of particles in the feasible region [16, 17].

Our paper arranged into three sections. Section II introduces the review deeply on distinguishing and detecting twitters spam. In Section III, the big fundamental truth of our work explained. Section IV contributes the fundamental analysis of streaming based twitter spam detection by using hybridization of Decision Tree (DT), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) combined form these hybrid are (PSG-DT). Section V provides the conclusion of the proposed work. With the brief statement, our endowment of this paper is,

1. We generate the fundamental truth for our research on detection of spam tweet. Our research produce the detail effects related to the data factors, including spam to non-spam, data size training and performance of our detection ratio.
2. Here we found transfer of continuous function is important feature for the performance of spam detection. 10 lightweight features are extracted from twitter API and detected streaming of spam tweet. Also the time consumption and variation found newly for this spam tweets.
3. Under different experimental settings the behavior of our model are reported and experimented the detection of spam tweets using the decision tree with combination of three classifiers.

2 Related Work

The awareness on streaming twitter spam is already got attention by the researchers. Distinguishing of twitter spam have studied by some researchers, and also proposed some important works for the detection of spam in twitter. For this, we study the previous related works by classifying into two categories: 1) Distinguish twitter spam and 2) Detecting twitter spam.

2.1 Distinguish Twitter Spam

To perceive the intended sense behind twitter spam, a detailed scrutiny has been carried out. In 2010, researchers examine 25 million URLs to detect spam and these URLs were taken from 200 million tweets posted by twitter users, then they found 8% of close URLs that is about 2 million spams were detected [3]. From their research they concluded email spam is less harmful than that of Twitter spam, they compared using CTR method which shows email spam has a less rate (0.0003–0.0006%) than twitter spam having high rate about 0.13%. The result from the implementation of blacklist has failed to stop the spam from the users who are affected.

Yang et al. proposed an analysis on ecosystem of cyber criminals and the relationship about spams by gathering information from criminal and supporter's community on twitter. They gathered 2060 criminal accounts and connection to this social website, which separated as small world and hub like model to connect many accounts to follow them. From this research Yang et.al created a conclusion algorithm for the criminal accounts to gather hidden spammers using the existing spammers [18, 19].

In 2011, Thomas et al. collected large number of dataset about 1.8 billion tweets for detecting spam characteristics which result shows 80 million were affected [5]. A group of five spammers' community was found and analyzed the behavior. Many esteemed online shopping websites like Amazon were affected, which become less distinct due to spam. From their research 77% spam accounts were detected and disabled when first tweet was posted and then 92% of accounts were suspended within 3 days. Due to this condition 89% of spammers accounts having connections to the user.

2.2 Detecting Twitter Spam

Many researchers have proposed machine learning algorithm based spam detection to distinguish spam and non spam. More important works have proposed [1, 2, 20–22], which used for report and content features like age of account, number of followers, tweet length, URL proportion, which used to characterize spammers and non-spammers. These features are effectively extracted and also it can be falsified easily. Accordingly, the work [23, 24] proposed is to avoid falsified feature; researchers used robust features to move which depend on social graph. Song et al. proposed a work to detect spam tweet by extracting connection and distance between sender and receiver in twitter [23]. Meanwhile in this [24], Yang et al. proposed a work based on graphical view of this social network to form robust features, which include ratio of bidirectional link, centrality, and factor for local clustering. These robust features proved to be the best characterized than that of earlier works. Due to very large social graph in twitter the collection of certain features would be consuming more time and resource. Also the incoming of tweets are in the form of stream, so it is unfeasible to gather those features.

Alternatively from [7, 25] have proposed a work to detect spam by simply depends on embedded URLs in twitter. In [25], some amount of URL features are used like path tokens, domain tokens and URL parameters, with including other features used from particular section of twitter website like DNS information, and domain information. In [7], redirected chains of URL attribute have been studied and further collected the URL features like chain length and number of different initial URLs for classifying spam by characterizing their ability. Although, the two works performed above is only for URL spam detection, which shown by recent works [7, 25]. Thomas et al. [25], proposed a model for detection of spam for each user such as language model and time posting model. The spammers use these accounts to create spam or spamming activities when it is compromised, it usually happens when the model is failed to do its job. From this method it only determines whether the user compromised or not, but does not fit to detect fraudulent accounts.

Several works proposed for the detection of spam when streaming spam tweets are analyzed. The concepts and ideas have developed and applied for the spam detection by some methods using Machine learning algorithms, which proposed for the evolution of streaming twitter spam, but still there is absence of performance which is going to be processed further. In this paper, we propose a combination of ML algorithm with Evolution algorithm to improve the performance and detection rate for streaming of spam tweets.

3 Streaming of Public Tweet Dataset from Twitter

Researchers proposed many algorithms for the detection of streaming spam tweets by collecting large dataset with fundamental truth to perform many actions. Here we use the combination of algorithms to perform the detection of spam tweets. Although streaming of datasets is not available publicly for our work, we use example tag as spammers instead of spam tweets which published from previous research [1, 2, 24]. As an outcome, we collect a large datasets from streaming twitter spam over 600 million tweets and prompt for fundamental truth. Also the dataset includes more than 6.5 million spam tweets and then we make this dataset available for future research. Further this section going to see the procedure to collect twitter datasets, fundamental truth, extracted features, and statistics of attribute.

3.1 Procedure to Collect Twitter Dataset

Twitter not allowed to fully accessing the dataset by using streaming twitter API but it granted 1% of his tweets to use for the research work. It would not allow approaching preserved accounts and direct messages due to its restrictions of company policies. The twitter streaming API collects public tweets with URLs [27] and it extracted using JSON format (see Fig. 2 for Tweet JSON example), each line of code represents an object which is simple and easy to be access. Several attributes are available in this streaming API of twitter like “hastags”, “URLs”, “retweets”, “Text”, “account generated time”, number of tweets”, “number of friends” [28]. Almost all the spam and unwanted messages contains URLs in twitter platform but in Twitter it is able to send spams and unsolicited messages without URLs [26]. During the research we found that some tweets not having URLs when collecting thousands of spam tweets manually. Spammer’s main goal is actually to use planned URLs which are needed by victims and allow them to use their sites in the names of phishing, scams and viruses downloading [29]. For our research we have collected 600 million tweets with URLs, which we restrict by not having tweets with URLs [30].

3.2 Fundamental Truth

Physical examination [1, 2, 20], and filtering of blacklists such as Google safe browsing, [3, 18, 24, 31, 32], are two methods where researchers used for generating fundamental truth. Due to time and resource consuming, a little amount of data could be trained and labeled by physical examination. The tweets can be labeled using human intelligent task (HIT) websites, but still not used for process due to its expensive cost is higher and the results are not as expected [33]. Unfortunately twitter API restricts to label the spam tweets because the large amount of data is processed.

From our dataset of 600 million tweets, we recognize 6.5 million malicious tweets which considered about 1% from overall dataset. For recognizing these tweets for our fundamental truth we used Trend Micro's WRS to locate the URL esteemed spam tweets. Since the protection rate of Web Reputation service is about 99.8%, the outcome of this service is reliable and performed for analysis.

3.3 Extracted Features

We label the spam tweets which are extracted by using Twitter's Public Streaming API, but it actually returns random public tweets which are not socially connected and unable to process a social graph from this data. Social graph based features like local clustering and between's centrality [24], which is not possible to extract and distance calculation could not be processed due to same reason [23]. Our research mainly focus on real time streaming of spam tweets detection which can be accessed and determine form the tweet which is preferred. Table 1 lists the 10 extracted features from our dataset which includes the representation of them [34, 35]. These features can be divided into two sub categories, first one is user-based features and second one is tweet-based features. The first feature like 'user', 'account age', are extracted from JSON object which is calculated using the date when tweets are collected minus date of the account when created. Other user-based features like 'no_followers', 'no_friends', 'no_favourites', 'no_lists' are directly obtained from structure of JSON object. Secondly tweet-based features includes 'retweets_followers', 'retweet_lists', 'no_chars', but little computation is needed for

Table 1 Extracted features and its representation

Name of features	Representation of features
no_followers	Number of followers for this twitter user
no_friends	Total number of friends following this twitter user
no_favorites	Number of received favorites by this twitter user
no_list	Number of added list by this twitter user
no_statuses	Number of tweets statuses by this user
no_retweet_followers	Number of retweets of followers
no_retweet_friends	Number of retweets of friends
no_retweet_favourites	Number of retweets of favourites
no_retweet_list	Number of retweets of list
no_chars	Number of characters per tweet

these chars which would be counted from the tweet text itself, and the other two features can be directly extracted [36].

3.4 Features Statistics

Here, the overall statistics of each feature are proposed to analyze their characteristics by using empirical cumulative distribution function. Each and every feature has been plotted by using the same function as shown if fig. As we can see from fig (a) that spammers involved more likely as non-spammers for number of favorites. Next the followers of the user shown in fig (b) have spammers with little low ratio. Fig (c) shows the graph of number of friend for user and their cumulative distribution which have the same ratio, but fig (f) has shown the retweets of followers count whose spammers are lesser than that of non-spammers. The important motivation of the spammers is to attract the user and motivate them to follow their links and connect with them.

Fig (i) shows the retweets of lists count, in which the spammer's lists are high in ratio than that of the non-spam. The user becomes victims as far as he has been retweeted by the spam lists. Fig (i) showing the number of characters of each tweet sent by the user and the equality of both the non-spammers and spammers are equal in range.

4 Streaming of Spam Tweets Detection Using Hybrid Algorithms

Here, the hybrid combination of algorithms is PSO, GA and DT, which perform an analysis of spam over streaming of tweets. We carry out evaluation on our dataset to detect the spam ratio. Also, we are going to perform an analysis of different datasets which are sampled between continuous and non-continuous method which are listed in Table 2.

From the above table we can see the ratio of spam to non-spam tweets of datasets I and II having 1:1 and from datasets III and IV having 1:19. From the earlier work, spam to non-spam ratio is 1:1 which is equally separated. Actually, in real world the Twitter has only 5% of spam tweet from all currently operating tweets [3]. There is lot difference between these evenly distributed datasets and Twitter samples which could not replace each other. Accordingly for this reason we choose dataset III and IV for simulation to real world framework because of its ratio 1:19.

As of now the whole datasets are obtained from 600 million tweets and separated as four types. However, by applying sampling method we represent the datasets into two

Table 2 Sampled datasets

Sample dataset	Methods	Amount of spam tweets	Amount of non-spam tweets
I	Continuous	5000	5000
II	Non-continuous	5000	5000
III	Continuous	5000	95,000
IV	Non-continuous	5000	95,000

```
{
          "text": "RT @PostGradProblem: In preparation for the NFL lockout, I will be spending
    twice as much time analyzing my fantasy baseball team during ...",
          "truncated": true,
          "in_reply_to_user_id": null,
          "in_reply_to_status_id": null,
          "favorited": false,
          "source": "<a href=\"http://twitter.com/\" rel=\"nofollow\">Twitter for iPhone</a>",

          "in_reply_to_screen_name": null,
          "in_reply_to_status_id_str": null,
          "id_str": "54691802283900928",
          "entities": {
                "user_mentions": [
                      {
                     "indices": [
                                 3,
                                 19
                           ],
                           "screen_name": "PostGradProblem",
                           "id_str": "271572434",
                           "name": "PostGradProblems",
                           "id": 271572434
                ],
                "urls": [ ],
                "hashtags": [ ]
          },
          "contributors": null,
          "retweeted": false,
          "in_reply_to_user_id_str": null,
          "place": null,
          "retweet_count": 4,
          "created_at": "Tue Mar 23 23:48:36 +0000 2017",
```

Fig. 3 Example of JSON objects from Twitter streaming API

groups: The first group is collection of datasets I and III from whole samples without conscious decision with continuous flow of tweets. And the second one is datasets II and IV with non-continuous flow. Alternatively the tweets sent were free of control from each other.

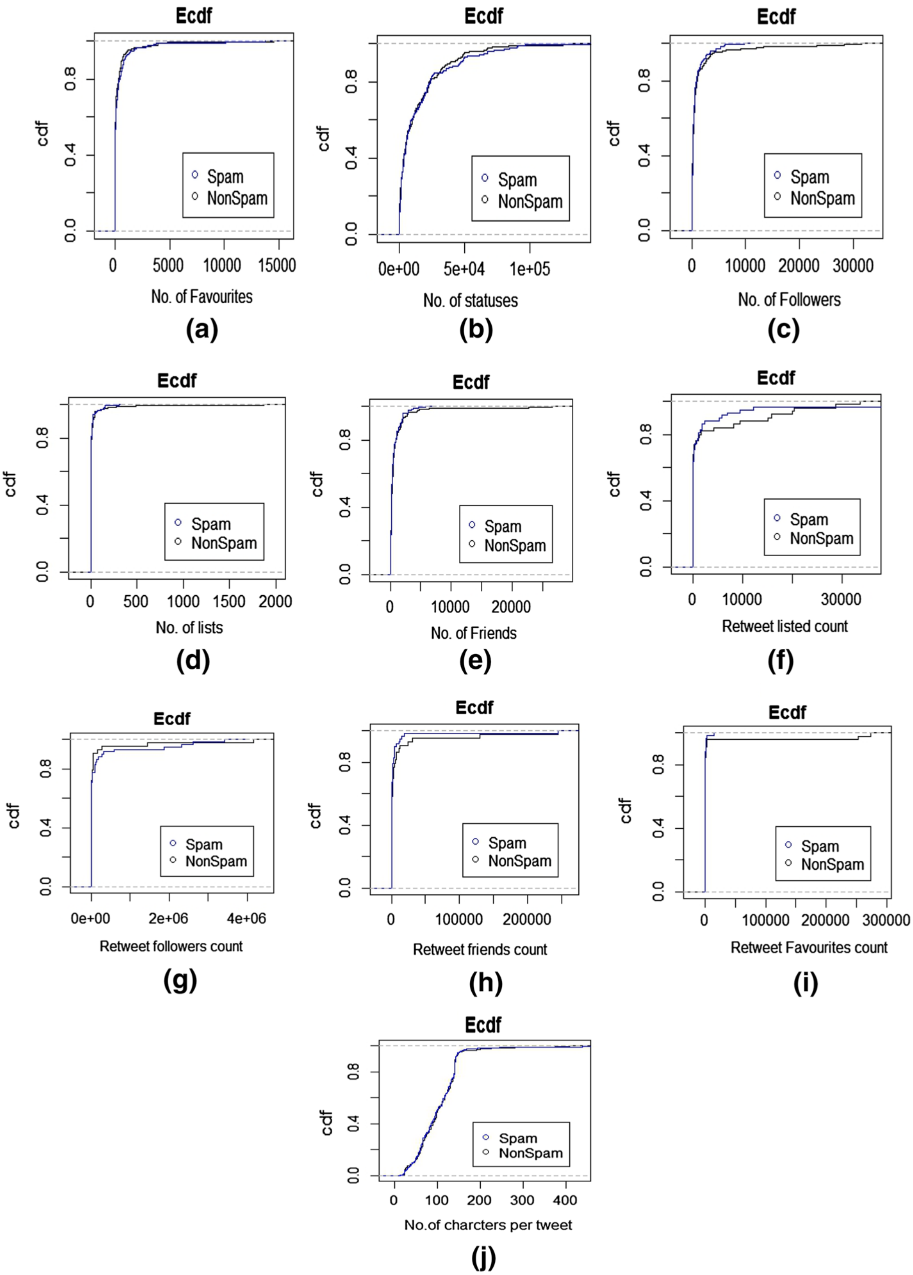


Fig. 4 Empirical cumulative distribution functions of features. **a** Number of user favorites. **b** Number of user followers. **c** Number of user friends. **d** Number of user statuses. **e** Number of user lists. **f** Retweets followers count. **g** Retweets friends count. **h** Retweets favorites count. **i** Retweets lists count. **j** Number of characters per tweet

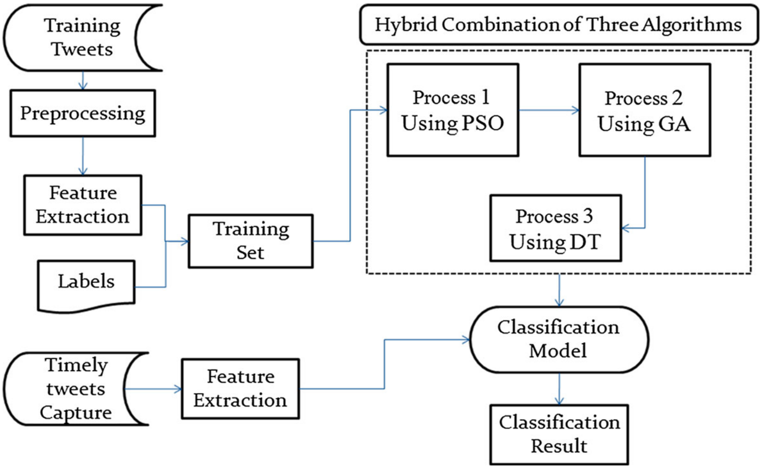


Fig. 5 Detection of twitter spam

4.1 Detection of Twitter Spam

Here our work proposed for detecting spam in twitter is processed using two algorithms namely genetic and decision tree. Figure 3 demonstrates the progress developed to detect streaming of spam using classifier and heuristic optimization techniques. The pre-labeled tweets which have taken before are trained by the classifiers which have the knowledge of developing general structure before performing the classification process. In order to predict the upcoming current tweet, the classification model should obtain some knowledge structure from the given datasets (Figs. 4, 5).

The method for detection of spam consist of mainly two process (1) Gaining Knowledge and (2) Classifying. Firstly the training tweets are extracted in the forms of vectors,

$$\vec{\mathbf{V}} = \{\mathrm{V}_1,\ \mathrm{V}_2 \ldots\ \mathrm{V}_n\}$$

Then using manual inspection approach we elaborate the class labels into spam and non-spam. After collecting the class labels and extracted features we mingle both data for training set in single sample. For further process, tweet can be represented as combination of one training tweet with the feature vector, and the result produce by these two connections is ($\vec{V}$, label), and training set vector can be given as

$$\overrightarrow{\mathbf{TS}} = \{(\vec{\mathbf{V}},\ \mathrm{label}_1),\ \{(\vec{\mathbf{V}}_2,\ \mathrm{label}_2),\ \{(\vec{\mathbf{V}}_\mathrm{n},\ \mathrm{label}_\mathrm{n})\}.$$

However this training set of data is the input for genetic algorithm operator, after computing GA process, the new population created would be used by classification model (decision tree) which then classifies spam to non-spam.

4.2 PSG-DT Algorithm

```
For each class in training set do
Construct numerical vector of each class
End for
Step 1: Let i=0, Ps=spam and S=new population;
Step 2: Initial population from training set is TS = {($V_1$, $label_1$), ($V_2$, $label_2$) ($V_3$, $label_3$)……
($V_n$, $label_n$)}
Step 3: Evaluate fitness for PSO
Step 4: obtaining particle best
Step 5: Finding global best of TS(i)
Step 6: Update velocity and position TS(i)
Step 7: Repeat step 4 to step 6 until stopping criteria met
Step 8: Evaluate Fitness function for TS
    If (i=0) then    i++
Step 9: While TS (i) = Positive Do
            Ps (i)←TS(i).selectPositive();
            Ps (i)←rep roduction(TS(i));
            Mutate (Ps (i));
            Evaluate (Ps (i));
Step 10: While TS(i) = Negative Do
            Ns(i) ← TS(i).selectNegative();
            Ns(i) ← reproduction (TS(i));
            Mutate (Ns(i));
            Evaluate (Ns(i));
Step 11: S← bulid_new_population_from (Ps(i), Ns(i))
Repeat (Step 9 until the maximum population criteria met)
Step 12: Return S to GenDecTree
Step 13: GenDecTree(S, Timely tweets)
Step 14: While $S \neq 0$ do
             Spam ← 0
            Non_Spam ← Null
            e = Entropy(attributes)
Step 15: for all attributes A in (S, Timely Tweets) do
            Gain ← Information gain (A,e)
            If Gain =True then
            Spam ← Gain
            End If
            If Gain = False then
            Non_Spam ← Gain
    End for
Step 16: Partition (Spam, Non_Spam)
 Until stopping criteria reached
 End
```

4.3 Performance Metrics

A standard of measurement was introduced which are widely used by the researchers for the analysis of performance for several approaches over spam detection.

1. Positives and Negatives: Here, T represents the tweet sent by the user and S denotes the spam category. In order to find whether the tweet belongs to spam class first we have to analyze the T (tweet) by sending it to GA operator and optimize the dataset to

create new population which further taken for the classification using decision tree algorithm and check whether it belongs S (spam category) or not. The action of classifiers can be analyzed using True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [37]. Given metrics can be classified into following:

a. The True positive performs actual tweet T of spam class S is member of that class S.
b. True Negative correctly analyzed that tweet T is not a member of spam class S.
c. False Positive wrongly suggested that tweet T is belong to class S.
d. False Negative tweet is member of spam class S but it wrongly suggests T not a member of class S.

The common relationship upon social network spam detection is given in Table 3.

Later on we calculate the True Positive Rate (TPR) and False Positive Rate (FPR) to measure the capacity of the classifiers for detecting spam. From [1] we come to a conclusion formula for this,

$$\text{(i)}\quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}\ + \mathrm{FN}}$$

$$\text{(ii)}\quad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{FN}}$$

2. Precision, Recall, and F-measure:
To analyze the variability of the per-class performance of the algorithm we calculate these three literatures.

a. Precision can be intended as ratio of tweets that exactly member of spam class S which recognized as class S.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

b. Recall can be calculated as ratio of tweets which is a member and exact category of spam class S to the complete number of users in class S.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

c. F-measure is the mixture of Precision and Recall, which acquire a large range of analysis of accuracy and average.

$$\text{F-measure} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Table 3 Analyzing metrics

Metrics	Prediction of Spam	Prediction of non-spam
Spam	True positive	False negative
Non-Spam	False positive	True negative

4.4 Spam to Non-Spam Collision

Here, the evaluation process of non-spam to spam ratio is analyzed using our proposed hybrid method on four sample datasets. The main motivation of our research work is to train our new classifier by using the datasets shown in Table 2. Hence, the trained classifier is used for spam detection. From [24], for the evaluation process further we analyzed TPR, FPR and F-measure using the classifiers. From Table 4, the performance of hybrid algorithms is compared with the proposed work which shows a high detection rate than that of other approaches. The performance evaluation for Dataset I for PSO-GA give 91.2% of true positive rate with 5.4% of false positive rate and overall F-measure gives 94.2%. And PSO-DT has 91.5% TPR with 5.9% FPR and finally 93.7% of accuracy is obtained. GA-DT has little less F-measure than of other two hybrid algorithms, but our proposed work PSG-DT give a high performance of 93.8% of True positive rate with 95.4% F-measure. Henceforth, Table 4 shows the overall performance results of each and every algorithm with our proposed work for all four dataset. The difference on performance is shown when the dataset is continuous as well as for non-continuous and amount of non-spam and spam taken are shown in Table 2.

As shown in Table 4, the analysis of our proposed approach is lesser when we apply for dataset III and IV which the amount spam to non-spam ratio differs and performance of each classifier getting less TPR and FPR also the F-measure. Because of the ratio between the non-spam and spam increases for the dataset III and IV the detection rate decreases and we can see from Table 5 the confusion matrix of the proposed work which defines the impact of both datasets. So as the precision decreases when the increase in amount of ratio. But, still our work gives better results than that of other machine learning classifiers which shown in Table 4.

Table 5 shows the confusion matrix of the proposed work, when the amount of data is increased with continuous and non-continuous dataset. Hence, the performance over proposed work has shown the detection of spam to non-spam ratio. From Fig. 6, the true positive rate of our classifiers is shown between discretization and non-discretization of Dataset I and III.

From Fig. 7, it has shown the false positive rate of dataset I and III which is non-discretized data gives less than the discretization of spam features. Even after the feature discretized, the proposed work has obtained deliberate results for FPR and has lesser than other algorithms. For F-measure, Fig. 8 shows the performance results on both datasets, in which PSO-DT has low accuracy for dataset III and alternative for other. AS of now, our work PSG-DT has high F-measure of above 90% for the discretization and lesser than 80% for non-discretization of features. From Fig. 6, 7 and 8, the performance of each hybrid classifiers is shown where PSO-DT is much lesser than other algorithms. But, the false positive rate of PSO-DT is less than the GA-DT due to its credibility of the algorithm. After the feature discretization, the Twitter spam detection can be improved by these hybrid classifiers.

4.5 Effect of Increasing Data for Training Process

So far, the evaluation of our hybrid techniques for all four Dataset has anlayzed and true positive rate, false positive rate and F-measure accuracy is evaluated. Now, the training size of Dataset I and III is increased for our further process and results are plotted and shown in Figs. 9 and 10. Here the data samples are increased from 100 to 1000 and

Table 4 Performance evaluation on four datasets

Unit%	Dataset I			Dataset II			Dataset III			Dataset IV		
Classifiers	TPR	FPR	F-measure	TPR	FPR	F-measure	TPR	FPR	F-measure	TPR	FPR	F-measure
PSG-DT	93.8	4.3	95.7	94.6	4.8	95.9	93.2	6.8	95.4	94.2	4.3	95.4
PSO-GA	92.2	5.4	94.2	92.2	5.7	94.1	91.6	7.1	94.2	91.7	5.3	93.1
PSO-DT	91.5	5.9	93.7	91.7	5.9	93.6	90.8	11.7	94.1	91.4	6.7	92.3
GA-DT	89.7	6.1	91.5	90.1	6.7	92.7	88.3	14.7	89.2	88.5	8.9	89.6

Table 5 Confusion matrix of PSG-DT

Classification→	Spam	Non-spam	Spam	Non-spam
Spam	4785	215	4653	347
Non-spam	7456	87,544	8765	86,235
	Dataset III		Dataset IV	

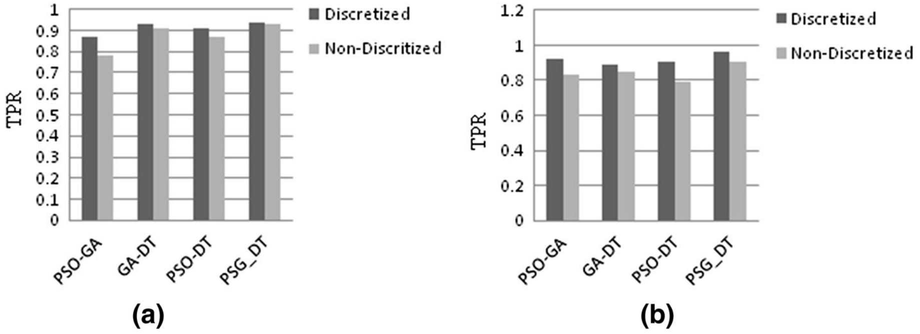


Fig. 6 True positive rate on spam: **a** dataset I and **b** dataset III

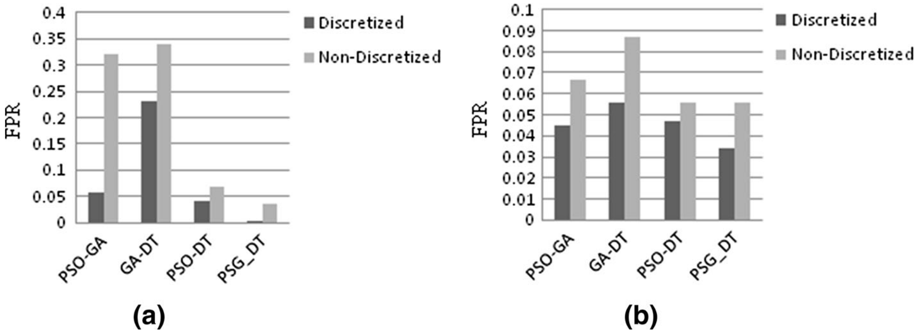


Fig. 7 False positive rate on spam: **a** dataset I and **b** dataset III

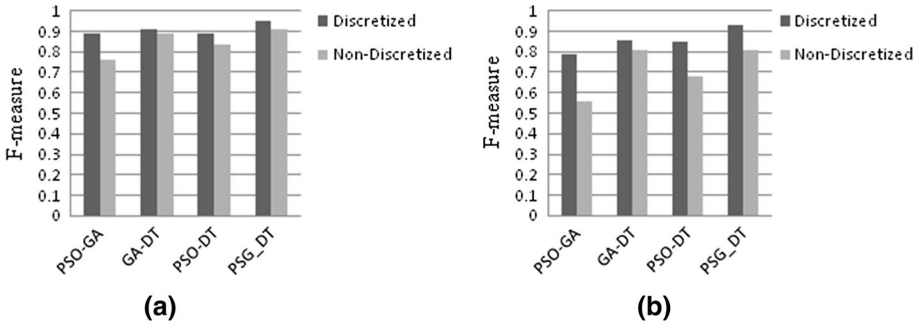


Fig. 8 F-measure on spam: **a** dataset I and **b** dataset III

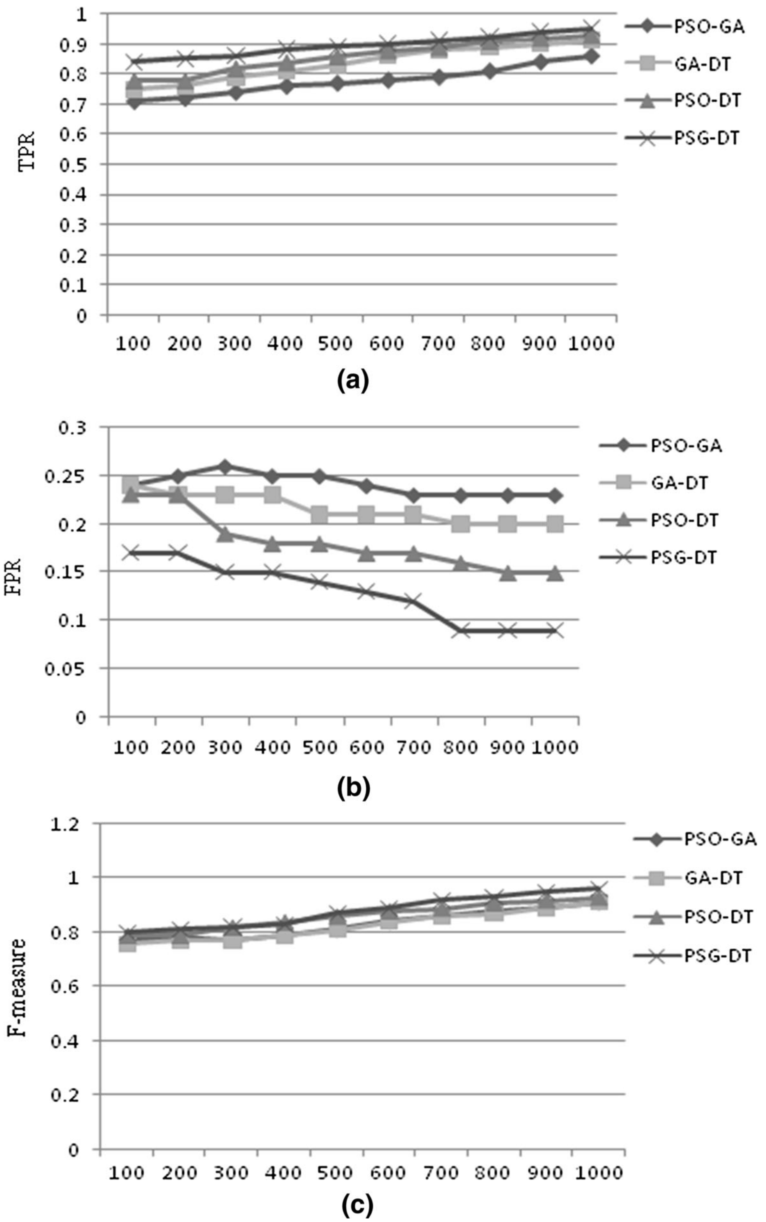


Fig. 9 Training size increased for dataset I: **a** TPR **b** FPR and **c** F-measure

analyzed the performance of each classifier. From Fig. 9, the dataset I is performed and increse the training samples which is in continuous flow and TP rate of each hybrid classifier is given and overall our proposed work PSG-DT has given better results which ranging from 80 to 90% than that of other algorithms and FP rate is lesser and PSO-GA has the higest range. After performing F-measure, the GA-DT has given lowest ratio and proposed algorithm given higher accuracy. Accordingly, for dataset III which is non-

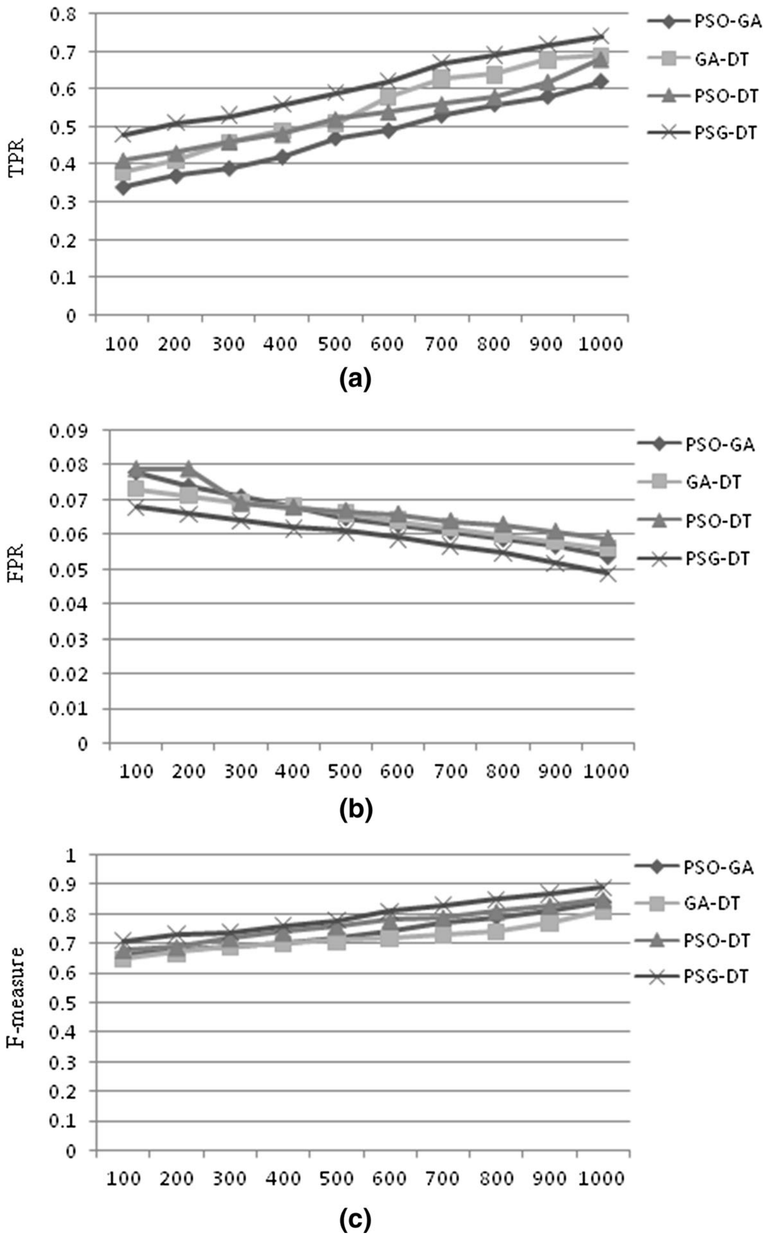


Fig. 10 Training size increased on dataset III: **a** TPR **b** FPR and **c** F-measure

continuous form and we increase the same ratio as of dataset I and perform TPR, FPR and F-measure for all hybrid algorithms and our proposed work has given better results than that of other classifiers with 70 to 90%. But still the performance is slightly differ from previous dataset due to the ratio of spam to non-spam.

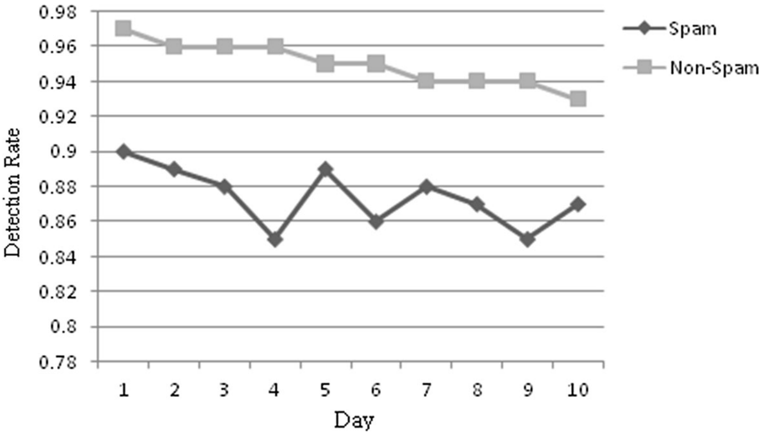


Fig. 11 Detection ratio of proposed work

4.6 Time Related Data analysis

After processing of our work with our training dataset we analyzed the performance evaluation for the each classifier. Now, the streaming of spam tweets is to be processed with 5 consecutive days and analyze the time related issues with each and every feature. By collecting a streaming data and perform the evaluation of proposed work with this new dataset which contains 100 k spam and 100 k non-spam tweets. From Fig. 11, the overall detection ratio of proposed work has proven to be best hybridization of all other classifiers, we only shown our PSG-DT results which predict the non-spam ratio from 90 to 97% and spam includes 80–90%. This experiment results are given from 10 consecutive days, each day the data is divided into half as spam and non-spam for the training and testing purposes. We extract the training data and train it with our proposed classifier with 10 k spam and 10 k non-spam tweets. As we can see the performance of our results slightly decrease from day 2 to 10. So as far the streaming of tweets collected for regularly 10 days can differs due to incoming of new type of spammers.

Therefore, timely tweets can be most challenging when compared to collected dataset due to daily millions of tweets are incoming with spammers finding new ways to send spam and our proposed work itself demanding to be developed much with some extending works in twitter dataset and increase the detection rate . So as to improve the classification accuracy, we need to analyze with different streaming of twitter data in daily basis.

5 Conclusion and Future Work

In this paper, the streaming of twitter spam is analyzed with our proposed work PSG-DT. Our work process first with collecting huge number of dataset consists of 600 million public tweets and we use 6.5 million tweets which we label it using Trend Micro’s Web Reputation system. After processing the label dataset we further classify our work with preprocessing step to analyze the missing values in our data. Later, feature extraction is done to extract particularly 10 features for the classification process to define non-spam and spam tweets after labeling the dataset. Furthermore, the empirical cumulative distribution of our features set is used to characterize each feature. So our proposed classifiers algorithms are performed using these feature set. Further we investigate with four different sampled dataset with various situations to simulate the spam detection. Hence, PSG-DT has proven to be the best classifier than that of other three classifiers which detection rate is increased for the collected dataset. GA-DT showed a less accuracy from overall evaluation. Unfortunately, our performance decreases for detection of spam when real-time imbalanced data is applied to our proposed work with other three classifiers.

From different days, the distribution of features changes and we can see the decrease in our classification accuracy. For future work, we would rather improve the performance of our proposed classifier for streaming of spam tweets which collected daily basis and analyze the detection rate. Hence, incoming new tweets could be problem as it comes in the form of streams. Furthermore, we extent this work in future.

References

1. Zhang, X., Zhu, S., & Liang, W. (2012). Detecting spam and promoting campaigns in the Twitter social network. In *2012 IEEE 12th international conference on data mining*. https://doi.org/10.1109/icdm.2012.28.
2. Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010) Detecting spammer on twitter. *Presented at the 7th annual collaboration electronic messaging anti-abuse spam conference, Redmond, WA, USA, Jul. 2010.*
3. Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010). @spam. In *Proceedings of the 17th ACM conference on Computer and communications security—CCS 10*. https://doi.org/10.1145/1866307.1866311.
4. Yardi, S., Romero, D., Schoenebeck, G., & Boyd, D. (2009). Detecting spam in a Twitter network. *First Monday*. https://doi.org/10.5210/fm.v15i1.2793.
5. Bohacik, J., Fuchs, A., & Benedikovic, M. (2017). Detecting compromised accounts on the Pokec online social network. In *2017 international conference on information and digital technologies (IDT)*. https://doi.org/10.1109/dt.2017.8024272.
6. Dasu, T., Krishnan, S., Venkatasubramanian, S., Yi, K. (2006) An information-theoretic approach to detecting changes inmulti-dimensional data streams. In *Proceedings of symposium on interface statistics and computer science applications*.
7. Lee, S., & Kim, J. (2013). WarningBird: A near real-time detection system for suspicious URLs in Twitter stream. *IEEE Transactions on Dependable and Secure Computing, 10*(3), 183–195. https://doi.org/10.1109/tdsc.2013.3.
8. Thonnard, O., Vervier, P., & Dacier, M. (2012). Spammers operations: A multifaceted strategic analysis. *Security and Communication Networks, 9*(4), 336–356. https://doi.org/10.1002/sec.640.
9. Chen, C., Zhang, J., Xie, Y., Xiang, Y., Zhou, W., Hassan, M. M., et al. (2015). A Performance evaluation of machine learning-based streaming spam tweets detection. *IEEE Transactions on Computational Social Systems, 2*(3), 65–76. https://doi.org/10.1109/tcss.2016.2516039.
10. Balan, E. V., Priyan, M. K., Gokulnath, C., & Devi, G. U. (2015). Fuzzy based intrusion detection systems in MANET. *Procedia Computer Science, 50,* 109–114.
11. Kumar, P. M., & Gandhi, U. D. (2017). Enhanced DTLS with CoAP-based authentication scheme for the internet of things in healthcare application. *The Journal of Supercomputing*. https://doi.org/10.1007/s11227-017-2169-5.
12. Hu, H., Chen, Y., & Tang, K. (2013). A novel decision-tree method for structured continuous-label classification. *IEEE Transactions on Cybernetics, 43*(6), 1734–1746. https://doi.org/10.1109/tsmcb.2012.2229269.
13. Manogaran, G., Thota, C., Lopez, D., Vijayakumar, V., Abbas, K. M., & Sundarsekar, R. (2017). Big data knowledge system in healthcare. In *Internet of things and big data technologies for next generation healthcare* (pp. 133–157). Springer.
14. Salehi, S., Selamat, A., & Bostanian, M. (2011). Enhanced genetic algorithm for spam detection in email. In *2011 IEEE 2nd international conference on software engineering and service science*. https://doi.org/10.1109/icsess.2011.5982390.

15. Rawal, B. S., Vijayakumar, V., Manogaran, G., Varatharajan, R., & Chilamkurti, N. (2018). Secure disintegration protocol for privacy preserving cloud storage. *Wireless Personal Communications*. https://doi.org/10.1007/s11277-018-5284-6.
16. Fukuyama, Y. (2008). Fundamentals of particle swarm optimization techniques. *Modern Heuristic Optimization Techniques*. https://doi.org/10.1002/9780470225868.ch4.
17. Modern Heuristic Optimization Techniques (2008). https://doi.org/10.1002/9780470225868.
18. Yang, C., Harkreader, R., Zhang, J., Shin, S., & Gu, G. (2012). Analyzing spammers social networks for fun and profit. In *Proceedings of the 21st international conference on World Wide Web—WWW 12*. https://doi.org/10.1145/2187836.2187847.
19. Balan, E. V., Priyan, M. K., & Devi, G. U. (2015, April). Hybrid architecture with misuse and anomaly detection techniques for wireless networks. In *2015 international conference on communications and signal processing (ICCSP)* (pp. 0185–0189). IEEE.
20. DON'T FOLLOW ME—Spam Detection in Twitter. (2010). *Proceedings of the international conference on security and cryptography*. https://doi.org/10.5220/0002996201420151.
21. Bhat, S. Y., & Abulaish, M. (2013). Community-based features for identifying spammers in online social networks. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining—ASONAM 13*. https://doi.org/10.1145/2492517.2492567.
22. Varatharajan, R., Vasanth, K., Gunasekaran, M., Priyan, M., & Gao, X. Z. (2017). An adaptive decision based kriging interpolation algorithm for the removal of high density salt and pepper noise in images. *Computers & Electrical Engineering*. https://doi.org/10.1016/j.compeleceng.2017.05.035.
23. Song, J., Lee, S., & Kim, J. (2011). Spam filtering in Twitter using sender-receiver relationship. *Lecture Notes in Computer Science Recent Advances in Intrusion Detection*. https://doi.org/10.1007/978-3-642-23644-0_16.
24. Varatharajan, R., Manogaran, G., Priyan, M. K., & Sundarasekar, R. (2017). Wearable sensor devices for early detection of Alzheimer disease using dynamic time warping algorithm. *Cluster Computing*. https://doi.org/10.1007/s10586-017-0977-2.
25. Thomas, K., Grier, C., Ma, J., Paxson, V., & Song, D. (2011). Design and evaluation of a real-time URL spam filtering service. In *2011 IEEE symposium on security and privacy*. https://doi.org/10.1109/sp.2011.25.
26. Thota, C., Sundarasekar, R., Manogaran, G., Varatharajan, R., & Priyan, M. K. (2018). Centralized fog computing security platform for IoT and cloud in healthcare system. In *Exploring the convergence of big data and the internet of things* (pp. 141–154). IGI Global.
27. Manogaran, G., Vijayakumar, V., Varatharajan, R., Kumar, P. M., Sundarasekar, R., & Hsu, C. H. (2017). Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. *Wireless Personal Communications*. https://doi.org/10.1007/s11277-017-5044-z.
28. Twitter. "Tweet structure" (2015). [Online] https://dev.twitter.com/docs/platform-objects/tweets.
29. Zhang, X., Zhu, S., & Liang, W. (2012). Detecting spam and promoting campaigns in the Twitter Social Network. In *2012 IEEE 12th international conference on data mining*. https://doi.org/10.1109/icdm.2012.28.
30. Chen, C., Zhang, J., Chen, X., Xiang, Y., & Zhou, W. (2015). 6 million spam tweets: A large ground truth for timely Twitter spam detection. In *2015 IEEE international conference on communications (ICC)*. https://doi.org/10.1109/icc.2015.7249453.
31. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., & Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. In *Proceedings of the 17th ACM conference on Computer and communications security—CCS 10*. https://doi.org/10.1145/1866307.1866396.
32. Devi, G. U., Balan, E. V., Priyan, M. K., & Gokulnath, C. (2015). Mutual authentication scheme for IoT application. *Indian Journal of Science and Technology*, *8*(26).
33. Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on world wide web—WWW 11*. https://doi.org/10.1145/1963405.1963500.
34. Kumar, P. M., & Gandhi, U. D. (2017). A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. *Computers & Electrical Engineering*. https://doi.org/10.1016/j.compeleceng.2017.09.001.
35. Varatharajan, R., Manogaran, G., & Priyan, M. K. (2017). A big data classification approach using LDA with an enhanced SVM method for ECG signals in cloud computing. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-017-5318-1.
36. Manogaran, G., Varatharajan, R., Lopez, D., Kumar, P. M., Sundarasekar, R., & Thota, C. (2017). A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring

and alerting system. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2017.10.045.

37. Zhang, J., Xiang, Y., Wang, Y., Zhou, W., Xiang, Y., & Guan, Y. (2013). Network traffic classification using correlation information. *IEEE Transactions on Parallel and Distributed Systems, 24*(1), 104–117. https://doi.org/10.1109/tpds.2012.98.

N. Senthil Murugan received his Bachelor's in Information Technology (with first class) from the Anna University, Chennai, Tamilnadu, India in 2011 and Master's in Computer and Communication Engineering (with first class) from the Anna University, Chennai, Tamilnadu, India in 2013, and currently working toward his Ph.D. in Information Technology in the VIT University, Vellore, Tamilnadu, India. His research interests include information security and social network security.

G. Usha Devi is working as an Associate Professor in the School of Information Technology and Engineering, Vellore Institute of Technology University. She received her Bachelor of Engineering in the University of Madras and Master of Engineering and Ph.D. degree from the Anna University. Her current research interests include big data analytics and network security. She has published number of international journals and conferences.