Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.020088 Published Online: 22 March 2022

ARTICLE

Email Filtering Using Hybrid Feature Selection Model

Adel Hamdan Mohammad[1,*], Sami Smadi[2] and Tariq Alwada'n[3]

[1]Computer Science Department, The World Islamic Sciences and Education University, Amman, Jordan

[2]Information System and Networks Department, The World Islamic Sciences and Education University, Amman, Jordan

[3]Network and Cybersecurity Department, Teesside University, Middlesbrough, UK

*Corresponding Author: Adel Hamdan Mohammad. Email: adel.hamdan@wise.edu.jo

Received: 03 November 2021 Accepted: 21 January 2022

ABSTRACT

Undoubtedly, spam is a serious problem, and the number of spam emails is increased rapidly. Besides, the massive number of spam emails prompts the need for spam detection techniques. Several methods and algorithms are used for spam filtering. Also, some emergent spam detection techniques use machine learning methods and feature extraction. Some methods and algorithms have been introduced for spam detecting and filtering. This research proposes two models for spam detection and feature selection. The first model is evaluated with the email spam classification dataset, which is based on reducing the number of keywords to its minimum. The results of this model are promising and highly acceptable. The second proposed model is based on creating features for spam detection as a first stage. Then, the number of features is reduced using three well-known metaheuristic algorithms at the second stage. The algorithms used in the second model are Artificial Bee Colony (ABC), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO), and these three algorithms are adapted to fit the proposed model. Also, the authors give it the names AABC, AACO, and APSO, respectively. The dataset used for the evaluation of this model is Enron. Finally, well-known criteria are used for the evaluation purposes of this model, such as true positive, false positive, false negative, precision, recall, and F-Measure. The outcomes of the second proposed model are highly significant compared to the first one.

KEYWORDS

Feature selection; artificial bee colony; ant colony optimization; particle swarm optimization; spam detection; emails filtering

1 Introduction

Emails are the most popular means of communication, and they can be considered a piece of life for most people. Emails are exposed to a lot of threats, such as spam [1]. Electronic spam is regarded as a serious problem and the most troublesome internet phenomenon challenging companies and individuals. In addition, spam causes traffic problems and bottlenecks that limit memory, power, and speed. Furthermore, spam is unwanted emails people and organizations send for several purposes, such as promotional, fraud, and other purposes. Also, it is the act of sending thousands or millions of emails to the recipient without their consent or approval. On the other hand, spam is being increasingly

used to distribute viruses, spyware, and different types of threats. Finally, spam is a severe threat, wastes internet traffic, contains malicious links, and is dangerous to our networks [2–5].

In the last decades, the internet has become one of the essential aspects of human life. Also, the internet is a crucial part of communication, education, and business. Emails are an electronic messaging method for communication, and it used to transfer messages from one user to another. In emails, junk is a primary concern, and spam emails are also called junk, unwanted, and unsolicited emails. At the present time, the number of spam messages has increased rapidly for several criteria such as advertisement, marketing, political emails, and other purposes [5–9].

Spam detection can be done in different methods and techniques such as content filter, header filter, whitelists, blacklists, machine learning, and others. Some of the methods used for emails classification are based on feature selection. Feature selection is a procedure for finding the minimum number of features from the original one. Besides, determining the minimum number of features is critical for the overall process success and failure [10–15].

Feature selection is mainly categorized into three methods: Filter method, wrapper method, and embedded method. The filter method is based on choosing the most significant features from the input before classification, such as correlation-based methods, information gain, and chi-square. The wrapper method is suitable for a dataset that contains fewer attributes. In wrapper methods, the results are better than the filter method, but it requires more computational. The embedded method is used to incorporate the filter method with the wrapper method. Examples of the embedded methods are weighted Naïve Bayes and artificial neural networks [6,16–25].

In this research, the authors focus on using metaheuristics and machine learning algorithms in developing an efficient feature selection for emails classification. The metaheuristics feature selection is essential for the success and the failure of any classification algorithm. Several feature selection methods are used in this research, such as Artificial Bee Colony (ABC), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO). Two models for emails classification and feature selection are used in this research. The first model is based on keywords reductions, while the second model is based on features reduction. **The contributions of this work can be summarized** in the following points:

- In model one, keywords are reduced from 3000 to 18 and 53 with promising results, and the emails spam classification dataset is used.
- In model two, the authors use adapted ABC (AABC), adapted ACO (AACO), and adapted PSO (APSO) for feature selection in a new manner. Also, the authors create 100 features for emails classification based on the subject and the body of the emails.
- In model two, each algorithm of AABC, AACO, and APSO are executed independently to reduce the number of features, and the most frequent features are selected for further experiments. Besides, in model two, the Enron spam dataset is used.

The rest of this paper is organized as follows: Section 2 talks about related studies. Section 3 demonstrates metaheuristic optimization and machine learning classification. Section 4 sheds light on the methodology used. Section 5 talks about results and discussions. Finally, this research conclusion is introduced in Section 6.

2 Related Studies

In this section, the authors will present several up-to-date research talks about emails classification and feature selection.

Bahgat et al. [26] proposed an efficient emails filtering approach based on semantic methods. The authors developed a system using the WordNet ontology and applied several semantic-based methods and similarity measures for decreasing a large number of extracted textual features. Also, the authors say that time complexities are reduced, and experimental is done on the standard benchmark Enron dataset. A comparative study for several classification algorithms is done, and the average accuracy is more than 90%. Alsmadi et al. [27] collected a large dataset of emails. Besides, several clustering methods were evaluated. The authors demonstrate that manual or supervised classification can be more reliable, and classification based on NGram is shown to be the best for a large text collection.

Ablel-Rheem et al. [28] used a data mining technique to classify spam emails. Besides, the dataset used in this research is UCI spam, and cross-validation is used for evaluation purposes in the training and the testing. Classifiers used in this research are Naïve Bayes, decision tree, ensemble booting, and ensemble hybrid boosting classifiers. Authors conclude that classification models using hybrid machine learning methods have an essential effect on spam detection. Matthew et al. [29] used phrases as a basic feature in emails classification. Many text classification methods were used in this research, such as Naïve Bayes, K-NN with TF-IDF weighting and resemblance. Investigation of this research includes the effect of phrase size, local and global sampling size, and neighborhood size. Authors conclude that public emails are easier to classify than private. Also, public emails used in this research are collected from different advertisements and newsletters. Govil et al. [30] proposed an algorithm to generate a dictionary and features and train them through a machine learning mechanism. Authors create a library named "stopwords" to remove all helping verbs from the content of the emails. Experiments were conducted using Naïve Bayes, and the dataset consists of 6000 emails. Aski et al. [31] described three machine learning algorithms for spam detection. The algorithms used are C4.5 decision tree, multilayer perceptron (MLP), and Naïve Bayes classifier. The dataset includes 750 emails and 750 spamand the authors demonstrate that the efficiency of MLP is better than the other models. Esmaeili et al. [19] proposed an email classification model using Naïve Bayes calssifier. Also, Implementation of feature selection is done with the help of ant colony optimization. The dataset used is collected from UCI Machine Repository, and the dataset contains 58 attributes. The evaluation was made based on accuracy, precision, recall, and f-measure.

Wu [32] presented a hybrid method of rule-based processing and a neural network for spam classification. The author utilizes spam behavior instead of keywords as features. Also, the rule-based is used to identify and digitize the spamming behavior from the headers and syslogs of emails. Besides, the author develops an enhanced back propagation neural network (BPNN) with a weighted learning strategy for the classification mechanism. On the other hand, the author mentions that the BPNN drawback is unstable time to converge. Also, the author demonstrates that It may not be possible to detect precisely all spam emails with a single technique. Hossam et al. [33] proposed an intelligent detection system based on the genetic algorithm and the random weight network for spam detection. Besides, the Authors develop an automatic identification capability that is embedded in the developed system. Datasets used are SpamAssassin, LingSpam, and CSDMC2010 Corpus. The results demonstrate the system can hit promising results. Ismaila et al. [34] proposed a model to improve the random generation of a detector in a negative selection algorithm (NSA) with the help of using stochastic distribution to model the data point using particle swarm optimization (PSO). The detector

generation process will be ended when the expected spam result is reached, and PSO is implemented at the random generation phase of NSA. Dataset used is obtained from spam base.

Yudongm et al. [35] proposed a spam detection model that focuses on reducing the false positive rate. The wrapper feature selection method is used to extract useful features. The authors use the C4.5 classifier model, and the dataset used contains 6000 emails. Also, the authors demonstrate that among seven meta-heuristic algorithms, the binary PSO with mutation operator (MBPSO) is better than the genetic algorithm (GA), restarted simulated annealing (RSA), particle swarm optimization (PSO), and binary particle swarm optimization (BPSO) in terms of classification performance. Also, the authors mention that using the wrap-based feature selection method can achieve high classification accuracy. Bilge et al. [36] proposed a detection method that combines artificial bee colony with a logistic regression classification model. Experiments were done on three publicly available datasets (Enron, CSDMC2010, and Turkish Email). Also, experiments are compared with support vector machine, logistic regression, and Naïve Bayes classifiers. The authors mentioned that the proposed model could handle high-dimensional data. Besides, the authors declared that one main limitation of the proposed method is its high computational complexity compared to other algorithms. Prilepok et al. [37] used two different algorithms for spam detection. The first algorithm is based on a Bayesian filter, while the second is based on particle swarm optimization. Preprocessing was done in this research, and numbers and words less than three characters were removed. The number of features is 300. The precision of ham detection is more than 99%, and spam is between 66% and 90%.

3 Metaheuristic Optimization and Machine Learning Classification

In this section, the authors will demonstrate several metaheuristic algorithms, such as artificial bee colony, ant colony optimization, and particle swarm algorithm. Also, this section will talk about random forest classifiers.

3.1 Artificial Bee Colony (ABC)

Artificial Bee Colony (ABC) is an optimization algorithm that was proposed by Karaboga in 2005 [38]. It is one of the swarm intelligence-based algorithms which mimic the behavior of the honeybee swarm. In the ABC model, the colony consists of three groups of bees: employee bees, onlookers, and scouts. In this model, it is assumed that there is only one artificial employed bee for each food source which means that the number of employed bees is equal to the number of food sources. Employed bees go to the source of the food and then come back to the hives. If the employed bee's source food is terminated, it becomes a scout. Onlooker bees watch the dances of employed bees and choose food sources depending on dances. Several types of research use ABC to get the optimal set of features for spam detection with significant performance [36,39–42]. ABC algorithm is adapted to fit the proposed model as follows:

Adapted ABC Algorithm

1- Initial Phase (**The dataset**)
   a. Allocate parameters to initial values (**Features**).
      Lower bound [1], upper bound [50,75,100], colony size [40,60,80,100], Max iteration [80,100]
   b. Arbitrarily generate solutions (**Feature**).
   c. Assess the solutions.
   d. Continue to employed bee phase.

(Continued)

2- Employed Bee Phase
  a. Create a candidate solution. (**Features**)
  b. Evaluate a new solution.
  c. Continue to onlooker bee phase (**search another feature**)
3- Onlooker Bee Phase
  a. Compute the probability of solutions. (**evaluate feature**)
  b. Pick solutions for each onlooker bee.
  c. Create the candidate solution.
  d. Assess a new solution (greedy selection).
  e. Continue to memorize the best solution.
4- Memorize the best solution.
5- Loop (While termination condition not satisfied) (Dataset is not empty)
  a. Is there another solution?
    i. If YES, Replace the best solution with a randomly generated solution (Scout Bee Phase), then Check stop criteria.
    ii. If STOP criteria, STOP or go back to the employed bee phase
  b. Is there another solution?
    i. If NO, If STOP criteria, STOP or go back to the employed bee phase
6- Output the best result (**Selected Features**)

3.2 Ant Colony Optimization (ACO)

Ant Colony Optimization (ACO) is a probabilistic technique for solving several computational problems. It is based on the behavior of ants seeking the best path between the colony and the source of the food [43,44]. the ACO technique was initially proposed by Marco Dorigo in 1992 in his Ph.D. thesis [45,46]. This technique consists of three steps: Firstly, each ant stochastically constructs a solution; later, the paths found by the different ants are compared, and finally, updating the pheromone levels on each edge occur. There are several types of research that use ACO in spam detection with highly acceptable performance. ACO algorithm is adapted to fit the proposed model as follows [19,39,40].

Adapted ACO Algorithm

1. Initialization
  a. Initialize pheromone evaporation [0.4,0.5,0.6], pheromone trail [1]. (**Features**)
2. Loop (While termination condition not satisfied) (**the dataset is not empty**)
  a. Ant generation. (**Feature**)
  b. Construct candidate solution.
  c. Pheromone evaporation.
  d. Update pheromone.
  e. Daemon action (if necessary).
3. Output the best result (**Selected Features**)

3.3 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) is an optimization algorithm that was developed by Kennedy and Eberhart developed it in 1995 [47,48]. It is inspired by the behavior of collective animals like birds and fish. Also, it aims to solve any problem by creating a population of candidate solutions

[49,50]. There are several types of research that use PSO in spam detection and classification. In PSO, each particle assesses its position based on the fitness function, and it is used to control its movement (velocity) in the existing search space. The velocity calculation considers the best position of the particles, which is called the personal best position (pBest), and the best position achieved by the neighbors (gBest). PSO particle positions are updated using the following equations. Also, PSO shows impressive results as an optimization algorithm and can be used with high-dimensional imbalanced data [51–55]. In addition, PSO particles positions are updated to fit the proposed model as the following Eqs. (1) and (2): [35,37,49,50].

$$Xi\,(t+1) = Xi\,(t) + Vi\,(t+1) \tag{1}$$

$$Vi\,(t+1) = W.Vi\,(t) + r1.c1.[pBest\,(i) - Xi\,(t)] + r2.c2.[gBest\,(i) - Xi\,(t)] \tag{2}$$

where:

X_i: The particle position i.

t: The iteration number.

V_i: The Velocity of particle i.

W: inertia weight.

r_1,r_2: random numbers between 0 and 1.

c_1,c_2: constant coefficients.

pBest: the current best position.

gBest: the current global best position of the particle's neighbors.

Adapted PSO Algorithm

1. Initialization
   a. initialize parameters (c1,c2,r1,r2) r1 and r2 [random numbers between 0 and 1; c1 and c2 [1,1.5,2,2.5,3,3.5])
2. Calculate fitness for each particle
3. Loop (While termination condition not satisfied) (the dataset is not empty)
   a. Evaluate fitness value; if better than pBest assign current as new pBest else; Keep previous value (PBest).
   b. Allocate best particles pBest to gBest.
   c. Compute velocity (See Eqs. (1) and (2)).
   d. Update velocity.
4. Output the best result (**Selected Features**)

3.4 Random Forest Classifier (RF)

Random forest (RF) is a classification algorithm consist of many decision trees. It is one of the most significant algorithms with high accuracy, and it was first introduced by Breiman in 2001 [56]. In RF, each tree is built based on different bootstrap samples drawn from the data to increase diversity. The number of selected features is less than the total number of available features in the dataset. One major advantage of RF is the speed and efficiency with large datasets [57–60].

4 Methodology

In this part, the authors will demonstrate the dataset used in this experiment. Also, feature selection and the proposed model are explained.

4.1 Datasets

Several datasets are available for evaluation purposes, such as PU, SpamBase, Enron spam, SpamAssassin, TREC, CCERT, and emails spam classification dataset [6,10]. In this paper, the emails spam classification dataset is used in model one, and It consists of 5172 emails [61]. The dataset of the emails spam classification consists of 5172 rows, each row for each email, and 3000 columns. The name of the emails has been set with numbers. The last column has the labels for prediction (1 for spam), and (0 for not spam), which means that the actual number of columns is 3000. The 3000 columns are the most common words in all the emails. The number of spam emails is 1500, and not spam is 3672, which means the percentage of spam emails is 29%, and the percentage of ham is 71%. Besides, in model two, the Enron spam dataset is used, and specifically, Enron6, which consists of 6000 emails, 4500 spam, 1500 legitimate, and all emails are text files [62].

4.2 Feature Selection

Feature selection is an important issue, and no doubt that feature selection and reduction play an essential role in any classification process. Creating features in emails classification can be done in several ways. In this research, features are created based on the subject and body of the emails. Features are constructed in similar ways as [59,60,63]. The number of features in this research is 100, 94 Features are based on symbol, word frequency, and the last six features are created based on other emails features. Part of the features used in this research is demonstrated in Table 1.

Table 1: List of features in model two

Features	Importance	Features	Importance
1. Symbol Freq !	0.402	90. World Freq Get	0.439
2. Symbol Freq &	0.394	91. World Freq Limited	0.321
3. Symbol Freq $	0.501	92. World Freq Now	0.568
4. Symbol Freq @	0.612	93. World Freq Action	0.621
5. Symbol Freq #	0.397	94. World Freq Urgent	0.612
...	...	...	...
20. World Freq Act	0.493	95. Number of words containing only letters	0.081
21. World Freq Action	0.485	96. Max of the character length of words	0.073
22. World Freq Apply	0.521	97. Number of words in all uppercase	0.0111
23. World Freq Buy	0.539	98. Number of words that are digits	0.0123
24. World Freq Call	0.502	99. Number of words containing letters and numbers	0.089
25. World Freq Click	0.569	100. Number of characters	0.071

4.3 The Proposed Models

The authors develop two models for spam detection. Model one is demonstrated in Fig. 1. While model two is shown in Fig. 2.

4.3.1 Model One

The first model is evaluated using the emails spam classification dataset. As mentioned previously, the dataset consists of 5172 emails (1500 spam, 3672 legitimate). 90% of the dataset is used for training, and the rest of the dataset is used for testing, see Fig. 1.

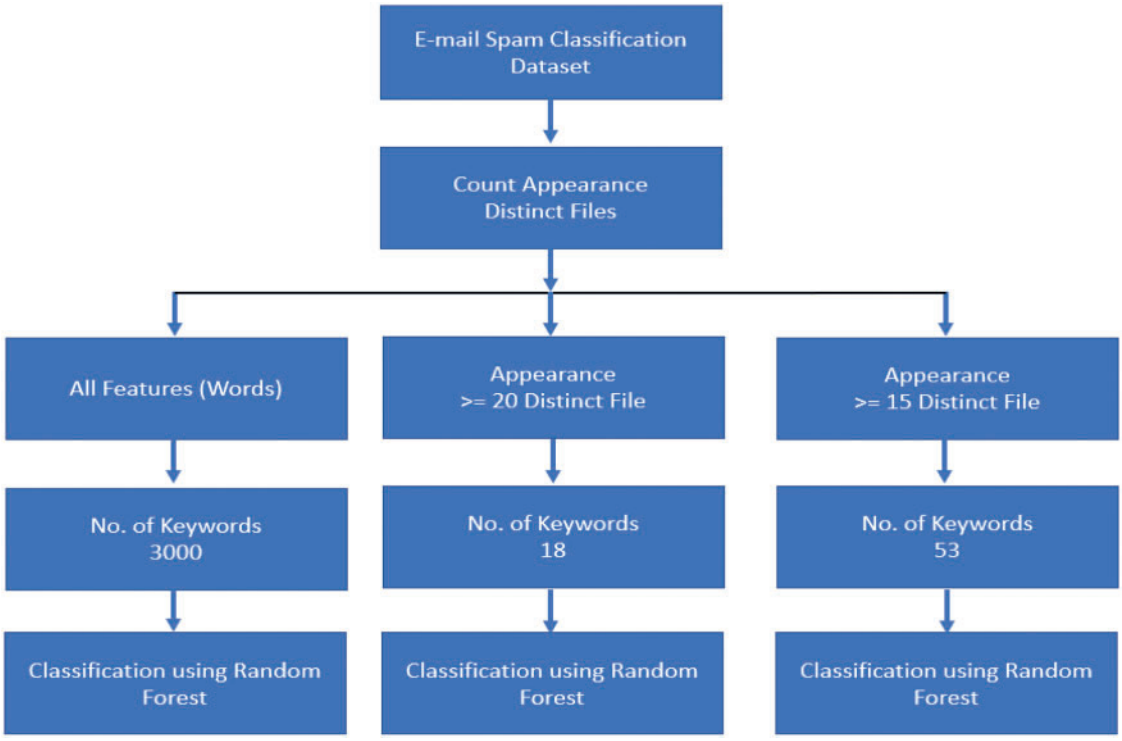

Figure 1: Model one (feature selection and classification using e-mail spam classification dataset)

Model one feature selection and classification are made as follows: the number of features is reduced based on the number of the appearance of any keyword (feature) in distinct files. Two experiments were done. The first one was with 18 keywords, for keywords appear in $>= 20$ distinct files. The second one was with 53 keywords, for keywords appear in $>= 15$ distinct files.

4.3.2 Model Two

The second model is evaluated using the Enron dataset. As mentioned previously, the Enron dataset consists of 6000 emails (4500 spam, 1500 legitimate). 90% of the dataset is used for training, and the rest of the dataset is used for testing, which means 4050 spam, 1350 legitimate for training, and 450 spam, 150 legitimate for the testing, see Fig. 2.

In this model, features are created based on the body and the subject of the emails. Prepossessing is done with this model. All words ≤ 3 characters are removed, and frequently symbols are considered. Experiments were done with only 100 features.

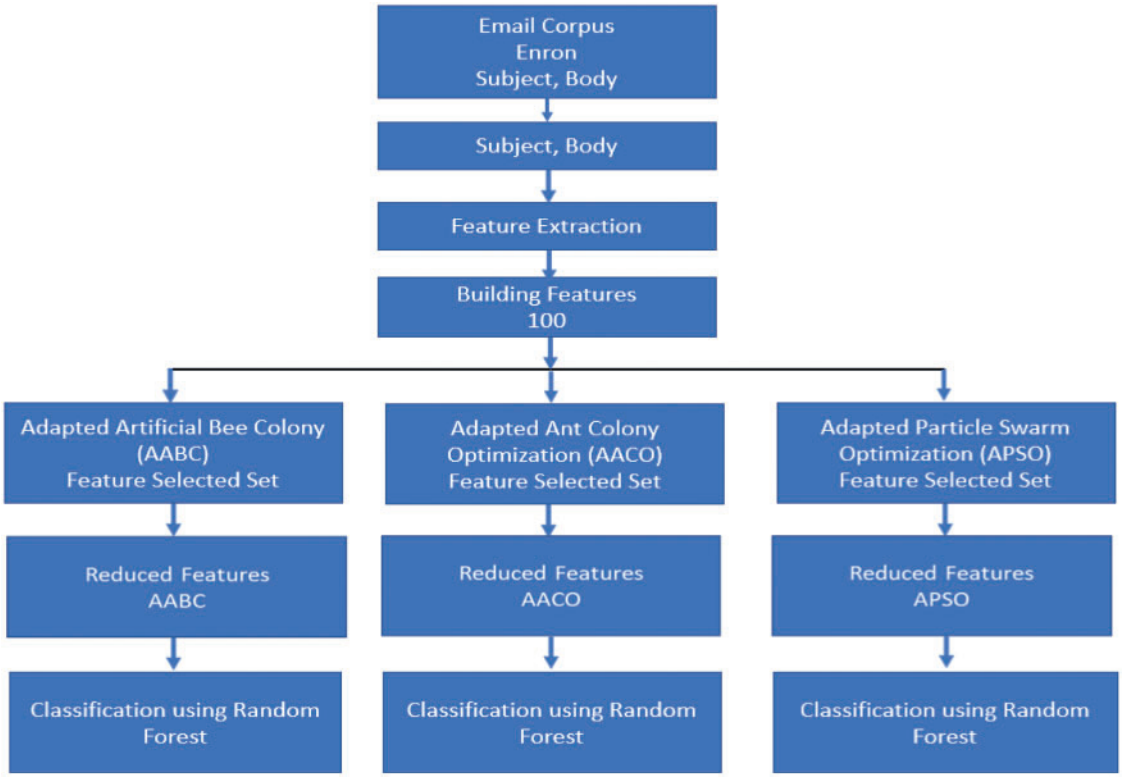

Figure 2: Model two (feature selection and classification using enron dataset)

5 Results and Discussions

The following section will demonstrate the evaluation metrics and the results related to this study. All experiments were done using Dell Machine, Intel(R), Core i7-CPU 1.8 GHz, installed memory (RAM) 16 GB, 64 Bit Operating System, Windows10. Besides. Finally, the Anaconda Python open-source is used.

The authors use standard criteria such as True Positive (TP), False Positive (FP), False Negative, Precision (P), Recall (R), and F-Measure. To demonstrate these regular criteria, please see Table 2 below and Eqs. (3)–(8).

Table 2: Confusion matrix

		Predicted	
		Normal	Attack
Actual	Normal	a (TP)	b (FN)
	Attack	c (FP)	d (TN)

TP (True Positive): The model correctly predicts the positive class.

FN (False Negative): The model incorrectly predicts the negative class.

FP (False Positive): The model incorrectly predicts the positive class.

TN (True Negative): The model correctly predicts the negative class.

Precision: The ratio of the number of correct decisions.

Recall: The ratio of total relevant results correctly classified.

F-measure: A single measure that balances precision and recall.

$$\text{TPR} = \frac{a}{a+b} \tag{3}$$

$$\text{FPR} = \frac{c}{c+d} \tag{4}$$

$$FNR = \frac{b}{a+b} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{6}$$

$$\text{Recall or (Sensitivity)} = \frac{TP}{TP+FN} \tag{7}$$

$$\text{F}-\text{measure} = \frac{2 * \text{Precision} * Recall}{(\text{Precision} + \text{Recall})} \tag{8}$$

5.1 Model One Experiments

Model one experiments were done as follows: the dataset used in this model was the emails spam classification. 90% of the dataset was used for training, and the rest was used for testing. The Anaconda Python open source was used as well.

The results of the proposed model one is shown in Table 2. The experiments were done using the RF classifier. In addition, the Anaconda Python open source was used at all stages. Table 3 demonstrates TP, FP, precision, recall, and the F-measure.

Table 3: Model one experiments using RF classifier

Appearance/Distinct files	No. Keywords	TP	FP	Precision	Recall	F-Measure
All files	3000	0.993	0.582	0.807	0.993	0.890
>= 15 Distinct files	53	0.950	0.257	0.901	0.950	0.925
>= 20 Distinct files	18	0.994	0.881	0.734	0.994	0.845

The FP rate result with 3000 keywords (all files) is 0.582, but it is 0.257 when using only 53 keywords that appear greater than or equal to 15 distinct files. Also, it is 0.881 when using only 18 keywords. The results show highly acceptable results with a reduced number of keywords. In addition, Fig. 1. demonstrates the TP, precision, recall, and F-measure results are highly accepted. The results with 53 keywords and greater than or equal to 15 distinct files are promising, see Fig. 3.

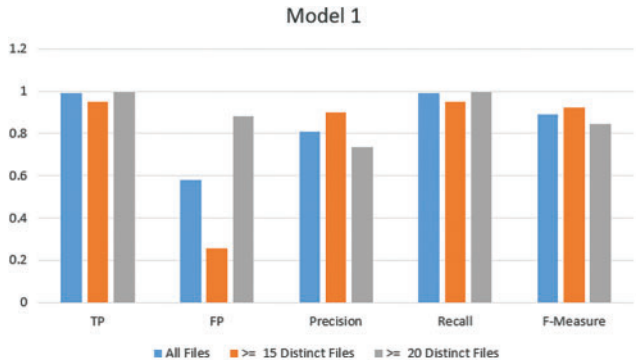

Figure 3: Model one results

5.2 Model Two Experiments

Model two experiments are done as follows: the dataset used is Enron. Also, experiments are done at two stages (training and testing). Training is done using 90% of the dataset, also preprocessing is done to eliminate the number of keywords and features. Finally, the Anaconda Python open source is used.

In model two, each feature selection method from AABC, AACO, and APSO are executed independently to select a subset from the 100 features created. The subset of the features is saved for further experiments. For experiments purposes, each algorithm is run independently 50 times, and the list of most repeated features is considered. Table 3 represents the number of features selected for each algorithm.

Table 4: Feature reduction

Feature selection algorithm	AABC	AACO	APSO
No. of features selected	47	54	63

The results of the proposed model two are shown in Table 3. Also, experiments are done using the RF classifier. Table 3 demonstrates TP, FP, precision, recall, and the F-Measure.

Model two experiments show highly acceptable results, especially with the APSO algorithm. As shown in Table 5, TP, FP, and FN results are promising with the reduced number of features. Figs. 4–9 show TP, FP, FN, precision, recall, and F-measure, respectively.

Table 5: Model two experiments using RF classifier

Feature selection algorithm	TP	FP	FN	Precision	Recall	F-measure
AABC	0.701	0.219	0.285	0.7620	0.7110	0.7356
AACO	0.852	0.159	0.258	0.8427	0.7676	0.8034
APSO	0.899	0.119	0.152	0.8831	0.8554	0.8690

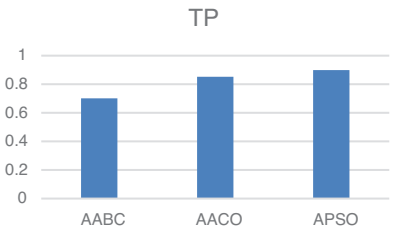

Figure 4: TP

Figure 5: FP

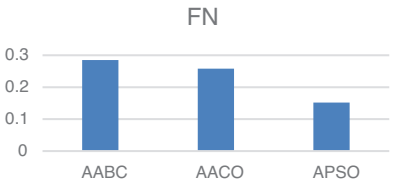

Figure 6: FN

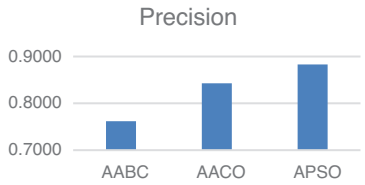

Figure 7: Precision

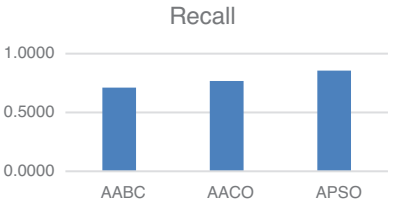

Figure 8: Recall

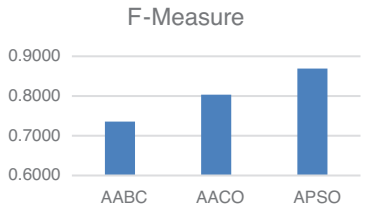

Figure 9: F-Measure

The results demonstrated in the figures above are highly acceptable. Also, Model two proposed using three well-known feature selections. The AABC, AACO, and APSO reduce the number of features from 100 to 47, 54, and 63.

F-measure values in both experiments are highly significant. Model one shows 92.5% F-measure with only 53 keywords. Also, the Best results in model two is demonstrated with APSO with 86.9%.

6 Conclusion

Spam is a serious problem that every user can face. This research proposes two models for spam detection and prevention. Model one for spam detection is evaluated with the email spam classification dataset. In this model, the number of keywords in the dataset is 3000. Also, this massive number of keywords is reduced to 18 and 53 by using two different criteria. Criteria used in model one is based on the number of the appearance of any keywords in distinct files. The results of model one are promising, and it shows the F-measure is 92.5% when selecting keywords that appear >= 15 distinct files. The second proposed model is based on using three emergent metaheuristic algorithms. The algorithms used in this model are Adapted Artificial Bee Colony (AABC), Adapted Ant Colony Optimization (AACO), and Adapted Particle Swarm Optimization (APSO). Several modifications are done on ABC, ACO, and PSO to fit the proposed model. In model two, 100 features are created based

on different criteria of email body and subject. Also, the number of features is reduced by using the AABC, AACO, and APSO algorithms to 47, 54, and 63, respectively. The limitation of this research is that it is evaluated with only two datasets. So it may be useful to perform the experiment on extra datasets. The results of the three mentioned algorithms are promising, but APSO shows the best results with 86.9% for F-Measure among the others. Future works for the authors could be evaluating other metaheuristic algorithms with different datasets.

Acknowledgement: Authors are very much thankful to the reviewers and Journal Authorities.

Funding Statement: The authors received no specific funding for this study.

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

References

1. Whittaker, S., Bellotti, V., Moody, P. (2005). Introduction to this special issue on revisiting and reinventing email. *Human-Computer Interaction, 20(1),* 1–9. DOI 10.1207/s15327051hci2001&2_1.
2. Adhav, K., Gawali, S., Murumkar, R. (2014). Survey on online spam review detection methods. *International Journal of Computer Science and Information Technologies, 5(6),* 7875–7876.
3. Crawford, M., Khoshgoftaar, T., Prusa, J., Richter, A., Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data, 2(23)*. DOI 10.1186/s40537-015-0029-9.
4. Jadhav, R., Gore, D. (2014). A New approach for identifying manipulated online reviews using decision tree. *International Journal of Computer Science and Information Technologies, 5(2),* 447–1450.
5. Hamdan, A., Abu-Zitar, R. (2011). Application of genetic ptimizoed artificial immune system and neural networks in spam detection. *Applied Soft Computing, 11(4),* 3827–3845. DOI 10.1016/j.asoc.2011.02.021.
6. Bhuiyan, H., Ashiquzzaman, A., Juthi, T., Biswas, S., Ara, J. (2018). A survey of existing email spam filtering methods considering machine learning techniques. *Global Journal of Computer Science and Technology: Software & Data Engineering, 18(2),* 21–29.
7. Hamdan, A., Alwada'n, T., Al-Momani, O. (2016). Arabic text categorization using support vector machine, naïve bayes and neural network. *GSTF Journal of Computing, 5(1),* 108–115.
8. Xue, B., Zhang, M., Browne, W., Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation, 20(4),* 606–626. DOI 10.1109/TEVC.2015.2504420.
9. Hamdan, A. (2021). Intrusion detection using a new hybrid feature selection model. *Intelligent Automation & Soft Computing, 30(1),* 65–80. DOI 10.32604/iasc.2021.016140.
10. Liu, Y., Wang, L., Shi, T., Li, J. (2022). Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. *Information Systems, 103(2),* 1–9. DOI 10.1016/j.is.2021.101865.
11. Salehi, S., Selamat, A., Kuca, K., Krejcar, O., Sabbah, T. (2017). Fuzzy granular classifier approach for spam detection. *Journal of Intelligent & Fuzzy Systems, 32(2),* 1355–1363. DOI 10.3233/JIFS-169133.
12. Masurah, M., Selamat, A. (2016). A New hybrid rough set and soft set parameter reduction method for spam E-mail classification task. *Knowledge Management and Acquisition for Intelligent Systems, PKAW, 22,* 18–30. DOI 10.1007/978-3-319-42706-5_2.
13. Choi, J., Jeon, C. (2021). Cost-based heterogeneous learning framework for real-time spam detection in social networks with expert decisions. *IEEE Access, 9,* 103573–103587. DOI 10.1109/ACCESS.2021.3098799.
14. Idris, I., Selamat, A., Nguyen, N., Omatu, S., Krejcar, O. et al. (2015). A combined negative selection algorithm-particle swarm optimization for an email spam detection system. *Engineering Applications of Artificial Intelligence, 39,* 33–44. DOI 10.1016/j.engappai.2014.11.001.

15. Mujtaba, G., Shuib, L., Raj, R., Majeed, N., Al-Garadi, M. (2017). Email classification research trends: Review and open issues. *IEEE Access, 5,* 9044–9064. DOI 10.1109/ACCESS.2017.2702187.
16. Sharma, V., Poriye, M., Kumar, V. (2017). Various classifiers with optimal feature selection for email spam filtering. *International Journal of Computer Science & Communication, 8(2),* 18–22. DOI 10.13140/RG.2.2.21621.06882.
17. Dagher, I., Antoun, R. (2017). Ham–spam filtering using kernel PCA. *International Journal of Computers and Communications, 11,* 38–44.
18. Kaur, H., Prince, E. (2017). Email spam detection using refined MLP with feature selection. *International Journal Modern Education and Computer Science, 9(9),* 42–52. DOI 10.5815/ijmecs.2017.09.05.
19. Esmaeili, M., Arjomandzadeh, A., Shams, R., Zahedi, M. (2017). An anti-spam system using naïve bayes method and feature selection methods. *International Journal of Computer Applications, 165(4),* 1–5. DOI 10.5120/ijca2017913842.
20. Shradhanj, A., Verma, T. (2017). Email spam detection and classification using SVM and feature extraction. *International Journal of Advance Research, Ideas and Innovations in Technology, 3(3),* 1491–1495.
21. Kumaresan, T., Saravanakumar, S., Balamurugan, R. (2017). Visual and textual features based email spam classification using S-cuckoo search and hybrid kernel support vector machine. *Cluster Computing,* 22, 33–46. DOI 10.1007/s10586-017-1615-8.
22. Zavvar, M., Rezaei, M., Garavand, S. (2016). Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine. *International Journal Modern Education Computer Science, 8(7),* 68–74. DOI 10.5815/ijmecs.2016.07.08.
23. Karthika, D., Visalakshi, P., Sankar, T. (2015). Improving email spam classification using ant colony optimization algorithm. *International Journal Computer Application,* 22–26.
24. Kumar, S., Arumugam, S. (2015). A probabilistic neural network based classification of spam mails using particle swarm optimization feature selection. *Middle East Journal of Scientific Research, 23(5),* 874–879. DOI 10.5829/idosi.mejsr.2015.23.05.22221.
25. Kalaibar, S., Razavi, S. (2014). Spam filtering by using genetic based feature selection. *International Journal of Computer Applications Technology and Research, 3(12),* 839–843. DOI 10.7753/IJCATR0312.1018.
26. Bahgat, M., Sherine, R., Walaa, G., Moawad, F. (2018). Efficient email classification approach based on semantic methods. *Ain Shams Engineering Journal, 9(4),* 3259–3269. DOI 10.1016/j.asej.2018.06.001.
27. Alsmadi, I., Alhami, I. (2015). Clustering and classification of email contents. *Journal of King Saud University-Computer and Information Sciences, 27(1),* 46–57. DOI 10.1016/j.jksuci.2014.03.014.
28. AblelRheem, D., Ibrahim, A., Shahreen, K., Almazroi, A., Ismail, M. (2020). Hybrid feature selection and ensemble learning method for spam email classification. *International Journal of Advanced Trends in Computer Science and Engineering, 9(1),* 217–223. DOI 10.30534/ijatcse/2020/3291.42020.
29. Matthew, C., Chung, K. (2009). Using phrases as features in email classification. *Journal of Systems and Software, 82(6),* 1036–1045. DOI 10.1016/j.jss.2009.01.013.
30. Govil, N., Agarwal, K., Bansal, A., Varshney, A. (2020). A Machine learning based spam detection mechanism. *Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 954–957. India.
31. Aski, A., Sourati, N. (2016). Proposed efficient algorithm to filter spam using machine learning techniques. *Pacific Science Review A: Natural Science and Engineering, 18(2),* 145–149. DOI 10.1016/j.psra.2016.09.017.
32. Wu, C. (2009). Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. *Expert Systems with Applications, 36(3),* 4321–4330. DOI 10.1016/j.eswa.2008.03.002.
33. Hossam, F., Al-Zoubi, A., Heidari, A., Aljarah, I., Mafarja et al. (2019). An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks. *Information Fusion, 48(4),* 67–83. DOI 10.1016/j.inffus.2018.08.002.
34. Ismaila, I., Selamat, A. (2014). Improved email spam detection model with negative selection algorithm and particle swarm optimization. *Applied Soft Computing, 22(1–2),* 11–27. DOI 10.1016/j.asoc.2014.05.002.

35. Yudongm, Z., Shuihua, W., Preetha, P., Genlin, J. (2014). Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. *Knowledge-Based Systems, 64(10),* 22–31. DOI 10.1016/j.knosys.2014.03.015.
36. Bilge, D., Bahriye, A. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. *Applied Soft Computing, 91,* 1–18. DOI 10.1016/j.asoc.2020.106229.
37. Prilepok, M., Jezowicz, T., Platos, J., Snasel, V. (2012). Spam detection using compression and PSO. *Fourth International Conference on Computational Aspects of Social Networks (CASoN)*, pp. 263–270. Brazil.
38. Karaboga, D. (2005). An idea based on honey bee swarm for numerical optimization. *Technical Report,* Erciyes University. https://abc.erciyes.edu.tr/pub/tr06_2005.pdf.
39. Akay, B., Karaboga, D. (2012). A modified artificial bee colony algorithm for real-parameter optimization. *Information Sciences, 192(3),* 120–142. DOI 10.1016/j.ins.2010.07.015.
40. Karaboga, D., Akay, B. (2009). A comparative study of artificial bee colony algorithm. *Applied Mathematics and Computation, 214(1),* 108–132. DOI 10.1016/j.amc.2009.03.090.
41. Singh, A., Chahal, N., Singh, S., Gupta, S. (2021). Spam detection using ANN and ABC Algorithm. *11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 164–168. India.
42. Sibel, A., Celal, O. (2018). Feature selection for classification with artificial bee colony programming. https://www.intechopen.com/chapters/68873.
43. Rungsawang, A., Taweesiriwate, A., Manaskasemsak, B. (2012). Spam host detection using ant colony optimization. In: *IT Convergence and Services, Lecture Notes in Electrical Engineering, 107,* pp. 13–21, DOI 10.1007/978-94-007-2598-0_2.
44. Manaskasemsak, B., Jiarpakdee, J., Rungsawang, A. (2014). Adaptive learning ant colony optimization for web spam detection. *International Conference on Computational Science and Its Applications, Computational Science and Its Applications*, pp. 8584, Portugal.
45. Colorni, A., Dorigo, M., Maniezzo, V. (1991). Distributed optimization by ant colonies. *European Conference on Artificial Life*, pp. 134–142. Paris, France: Elsevier Publishing.
46. Dorigo, M. (1992). *Optimization, learning and natural algorithms (Ph.D. Thesis)*. Italy: Politecnico di Milano.
47. Kennedy, J., Eberhart, R. (1995). Particle swarm optimization. *International Conference on Neural Networks, 4,* 1942–1948. DOI 10.1109/ICNN.1995.488968.
48. Eberhart, R., Kennedy, J. (1995). A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, pp. 39–43. Nagoya, Japan.
49. Kashif, I., Zainal, S. (2012). An improved particle swarm optimization (PSO)-based MPPT for PV with reduced steady-state oscillation. *IEEE Transactions on Power Electronics, 27(8),* 3627–3638. DOI 10.1109/TPEL.2012.2185713.
50. Marini, F., Walczak, B. (2015). Particle swarm optimization (PSO). A tutorial. *Chemometrics and Intelligent Laboratory Systems, 149,* 153–165. DOI 10.1016/j.chemolab.2015.08.020.
51. Zhang, Y., Wang, Y., Gong, D., Sun, X. (2021). Clustering-guided particle swarm feature selection algorithm for high-dimensional imbalanced data with missing values. *IEEE Transactions on Evolutionary Computation.* DOI 10.1109/TEVC.2021.3106975.
52. Song, X., Zhang, Y., Gong, D., Gao, X. (2021). A Fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for High-Dimensional Data. *IEEE Transactions on Cybernetics*. pp. 1–14. DOI 10.1109/TCYB.2021.3061152.
53. Song, X., Zhang, Y., Guo, Y., Sun, X., Wang, Y. (2020). Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data. *IEEE Transactions on Evolutionary Computation, 24(5),* 882–895. DOI 10.1109/TEVC.2020.2968743.
54. Hu, Y., Zhang, Y., Gong, D. (2021). Multiobjective particle swarm optimization for feature selection with fuzzy cost. *IEEE Transactions on Cybernetics, 51(2),* 874–888. DOI 10.1109/TCYB.2020.3015756.

55. Xue, Y., Xue, B., Zhang, M. (2019). Self-adaptive particle swarm optimization for large-scale feature selection in classification. *ACM Transactions on Knowledge Discovery from Data, 13(5),* 1–27. DOI 10.1145/3340848.
56. Breiman, L. (2001). Random forests. *Machine Learning, 45(1),* 5–32. DOI 10.1023/A:1010933404324.
57. Guo, L., Chehata, N., Mallet, C., Boukir, S. (2011). Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. *Journal of Photogrammetry and Remote Sensing, 66(1),* 56–66. DOI 10.1016/j.isprsjprs.2010.08.007.
58. Akinyelu, A., Adewumi, A. (2016). Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics, 6*. DOI 10.1155/2014/425731.
59. Faris, H., Aljarah, I., Al-Shboul, B. (2016). A Hybrid approach based on particle swarm optimization and random forests for email spam filtering. *8th International Conference on Computational Collective Intelligence*. Greece.
60. Alqatawna, J., Faris, H., Jaradat, K., Al-Zewairi, M., Adwan, O. (2015). Improving knowledge based spam detection methods: The effect of malicious related features in imbalance data distribution. *International Journal of Communications, Network and System Sciences, 8(5),* 118–129. DOI 10.4236/ijcns.2015.85014.
61. Email Spam Classification Dataset CSV. https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv.
62. Enron-Spam Dataset. http://nlp.cs.aueb.gr/software_and_datasets/Enron-Spam/index.html.
63. Khoi-Nguyen, T., Alazab, M. (2013). Towards a feature rich model for predicting spam emails containing malicious attachments and URLs. *Eleventh Australasian Data Mining Conference*, pp. 161–171. Canberra, Australia.