E-Mail Spam Filtering Through Feature Selection Using Enriched Firefly Optimization Algorithm

T. Poonkodi[a], Dr.S. Sukumaran[b]

[a]Ph.D Research Scholar, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, India. E-mail: ponrohit.0707@gmail.com
[b]Associate Professor, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, India.

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: E-mail is the most common method of communication because due to its ability to obtain, the rapid modification of messages and low cost of distribution. Spam causes traffic issues and bottlenecks that limit the amount of memory and bandwidth, power and computing speed. For data filtering, various approaches exist that automatically detect and suppress these indefensible messages. A methodology based on Sine- Cosine Algorithm (SCA) introduced which address the problem of space and time complexities are increased in E-Mail spam detection. In this method, WordNet optimized semantic ontology applies different methods based on semantics and similarity measures to reduce the large number of extracted textual features. This paper proposed the Enriched Firefly Optimization Algorithm (EFOA) method effectively selecting suitable features from an upper dimensional space using the fitness function. Once the best feature space is identified through EFOA, the spam classification is done using ANN. Intially, E-mail spam dataset is preprocessed, then the extracted textual features are Semantic-based reduction and Features weights updated using optimized semantic WordNet. The results obtained showed that the ANN classifier after selection of features using EFOA was able to classify e-mails as spam and non-spam. This EFOA demonstrates that the proposed method has led to a remarkable improvement compared to the SCA methods.

Keywords: E-mail, WordNet, EFOA Algorithm, Features Weight, Semantic based Reduction.

1. Introduction

With the growth in number of Internet users, e-mail has become the most widely used communication mechanism. Over the past few years, the increased use of emails has led to the emergence and aggravation of the problems caused by spam [1]. E-mails have maintained business communications leadership and continue to be a prerequisite for other electronic communications and transactions. The use of e-mails has led to a noticeable improvement in group communications, the impact of which is seen in growth of enterprises worldwide [2].

People use it for illegal and infernal purposes, phishing and fraud. Sending malicious link via spams that can damage our system and may also search your system. A spammer may collect the name of the individual who has a specific email address and include that name in the greeting of the message [3]. So, it is necessary to identify these spam mails which are frauds using ANN techniques.

The enriched firefly optimization algorithm (EFOA) [4-6] is a meta-heuristic algorithm. It is based on the communicating behavior of tropical fireflies. There are two important issues in the EFOA that involve changes in light intensity and formulation of attractiveness. The attractiveness of a firefly simply depends on its luminosity, but since attractiveness decreases with the distance between two fireflies, it seems that lower intensities involve less attractiveness. Even so, the EFOA still has a good capability. EFOA is also deficient, and will inevitably fall into local optimality, but its simple structure means that improving the algorithm has great potential. Address the shortcomings of EFOA by adding mechanisms to make it more effective.

The proposed approach Enriched Firefly Optimization Algorithm (EFOA) includes various components for select the optimal feature size to filter the E-mails using ANN classifier method into two classes: Spam and Non spam. The rest of the article is organized as follows: Section II presents a literature review related to the earlier spam detection techniques. Section III outlines the spam detection approach proposed EFOA. Section IV presents the performance analysis for the EFOA methodology. Finally, Section V concludes the whole discussion.

2. Related Works

Li et al., [7] proposed three different environments that are the Research Institute, the University and the business corporation in terms of their users. Five supervised basic machine learning classifiers were managed:

Naive Bayes, J48, IBK, Radial Basis Function Network (RBF-Network) and Library for Support Vector Machines (Lib-SVM). The result of the classification outcome indicates that the decision tree and support vector machines can produce better results than the other classifiers involved in this study [5].

Mallik et al., [8] proposed text parsing in the field of spam filtering to parse text that is embedded in junk mail. The Naıve Bayesian (NB) classification algorithm is used to construct the template and the R tool is used for the pre-processing step. This method identified the most commonly used topics more unused topics in spam emails. This approach could be used with various algorithms in order to obtain best results. It could also be utilized with hybrid algorithms to get the best results. Moreover, the orange software could be used to find out the outcome of each algorithm in a short period of time. After that, it might be developed into system of actual environment and organizational system.

Zhang et al., [9] proposed an automated detection approach specific to Chinese e-business websites by using the URL and functionality specific content of the website. Four machine learning classifiers were used, including RF, Sequential Minimum Optimization (SMO), logistic regression, and Naïve Bays (NB), and their results were evaluated using Chi-square statistics.

Laorden et al., [10] proposed the importance of finding anomalies discovery in UBE filtering to reduce the requirement for classification UBEs. Their work again analyzes an anomaly-based UBE screening approach that uses a data minimization approach that reduces pre-processing while maintaining information on the relevance of email messages to the nature of email. More recently, many task aimed at studying the suitability of various machine learning approaches including K-Nearest Neighbors (KNN), SVM, NB, neural networks, and others, to spam and malware email filtering, due to the ability of such viewpoint to learn, adapt, and generalize.

Awad & Foqaha [11] proposed a hybrid algorithm to optimize the rbf neural network and the particle swarm(HC-RBFPSO) for the classification spam emails. They used the particle swarm optimization algorithm to enhance the parameters of Radial Basis Function Neural Networks (RBFNN) based on PSO's scalable heuristic research. They split the dataset of spambase into 70% training set and 30% testing set. The experiments are measured using a different number of coverages from 10 to 50. The accuracy achived was 91:4% for the set of tests which concluded that the hybrid approach performed well compared to other algorithms tested on the same dataset.

Vyas et al., [12] offered alternative classification techniques using WEKA to filter spam mails. The technique of Naive Bayes demonstrated the apparent accuracy and least time among others. This paper provides a comparative review of all procedures in terms of accuracy and as well as the busy time is generated.

3. Proposed Methodology

EFOA based spam detection method is described in detail to address the problem of space and time complexities are increased and improve the E-Mail spam filtering. A critical step in the spam filtering of E-Mail is to find good feature selection and representation. The optimized semantic WordNet ontology is introduced for reducing the extracted textual feature and representation of updating weighted feature. It enhances the accuracy of spam filtering with high dimensionality of E-mail. Figure 1 shows the architecture of EFOA.

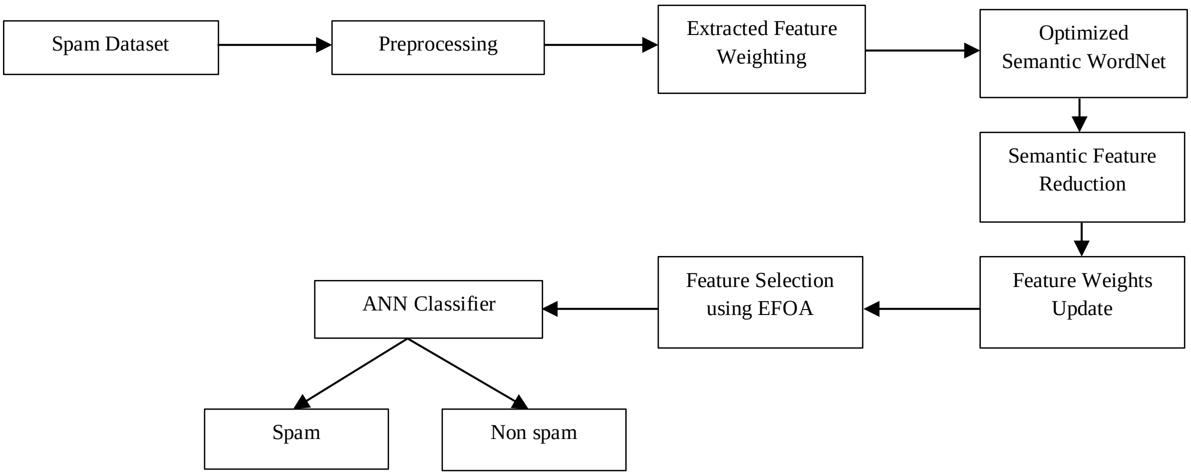
Spam Dataset
Preprocessing
Extracted Feature Weighting
Optimized Semantic WordNet
Semantic Feature Reduction
Feature Weights Update
Feature Selection using EFOA
ANN Classifier
Spam
Non spam

Figure 1. Architecture of EFOA based ESF

3.1 Preprocessing

In pre-processing, tokens are extracted and the nonrelevant tokens such as numbers and symbols are elliminated. The tokens are deleted from the body and subject matter of the Email. Following this, the stop words are removed. For more data cleansing, refer to WordNet as the Natural language Processing tool. WordNet is a optimize semantic network of words and there synonyms, antonyms, hyponyms, hypernyms, meronym and many more relations between words, which arrange English words into collection of synonyms called optsynsets [13]. Once an electronic document has been pre-processed, electronic messages may be represented by:

$$d_i = [(t_1, w_1), (t_1, w_1) \dots, (t_n, w_n)] \qquad (3.1)$$

where, each term ‘t’ is regarded as a feature that has a matching weight ‘w’ in a given outcome.

3.2. Feature Extraction and Calculate Feature Weighting

The weight of the retrieved feature, in which every term ‘t’ is weighted by a weight ‘w’ using the reverse document frequency and term frequency (TF-IDF) method. The frequency of the term indicates how many times the term ‘t’appears in the Email document ‘d’ as indicated in the Eq. (3.2).

$$tf(t,d) = (f_d(t))/\max[f_d(t)]) \qquad (3.2)$$

when $f_d(t)$ is the frequency for the title ‘t’ in e-mail. It measures the rarity of a certain term throught the document by means of equation (3.3).

$$IDF(t) = \log(N/df_t) \qquad (3.3)$$

where df_t is a number of e-mails with heading ‘t’, and ‘N’ represent the total number of e-mails. Lastly, the TF-IDF is calculated following the multiplication of Eqs. (3.2) and (3.3):

$$W = tf(t,d)IDF(t) \qquad (3.4)$$

3.3 Feature Reduction using Semantic Approach

In this process, synonyms for each feature are extracted, and then extracted features are replaced by their synonym ensemble concepts. In addition to extracting synonyms from each term, the hypernymous/hyponymic relationships are considered through the WordNet semantic optimization [21,13]. Word-to-word similarity is measured through various semantic similarity measures.

3.3.1 Semantic-based Reduction

The optimized semantic WordNet used as a method that stores the different forms of a word like such as English names, adjectives, verbs and adverbs to all synonyms. These synonyms are referred to as optsynsets, which are related to each other by semantic relationships. In semantic-based reduction, synonyms set of each term in the Email are used to group the terms that have common synonyms. The hyponymic relationships refers to ‘‘is a kind of” or ‘‘is a”, which connect new general optsynsets to specific ones, while the hypernymic link represents the converse of the hyponymic relation. After applying semantic relationships, various optimized semantic similarity measurements are applied to increase the rate of reduction rate for the features.

Path based Measurements

Steps based on path [20] are based on the length of path between two concepts. Three versions of similarity measure versions are tested for their performances that are path length measurement, WUP measurement and LCH measurement.

Path Length computes the semantic similarity of a concept pair by counting the number of nodes along the shortest path among concepts in WordNet’s ’is-a’ hierarchies. The path-like score is converse correlated with the number of nodes along the shortest path between the two words. Therefore, the equivalent metric equation is as follows:

$$PATH(t_1, t_2) = \frac{1}{length(t1,t2)} \qquad (3.5)$$

where t_1 and t_2 are either terms.

WUP measurement calculate similarity by examine the depths of the both terms in optimized semantic WordNet with the depth of the least common subsume (LCS) as shown the given equation:

$$WUP(t_1, t_2) = \frac{2\left(depth\left(LCS(t_1, t_2)\right)\right)}{depth(t_1)+depth(t_2)} \qquad (3.6)$$

Where the lowest common subsume (LCS) is the most specific common ancestor between two optsynsets $(t_{1,}t_2)$.

Leacock and Chodorow measure (LCH) obtain the shortest path-length among two concepts, and according to the given equation then scales the outcome value by the maximum depth found in the ‘‘is–a” hierarchy.

$$LCH(t_1, t_2) = -log\,\frac{length(t_1,t_2)}{2Max(depth)} \qquad (3.7)$$

The information Content (IC) provides an indication of the specific nature of the concept[18]. An IC measurements is the Resnik measurement which calculates the information content of the least common subsume (LCS) of the two terms, using the given equation:

$$Rensik\left(t_{1,}t_2\right) = IC(LCS\left(t_{1,}t_2\right)) \qquad (3.8)$$

The IC for a term (t) is defined as:

$$IC(t) = -logP(t) \qquad (3.9)$$

when P(t) the probability of a term (t) in a given messages with (N) separate terms.

$$P(t) = \frac{frequency(t)}{N} \qquad (3.10)$$

Relationship measurement: A different type of relationship measurement has been applied, namely HSO (Hirst and St-Onge)[16]. The HSO measurement estimates relation between the search terms by the journey distance between the nodes. The number of rework in the direction of the path linking two terms and the enable it of the path. The HSO function reads as follows:

$$HSO\left(t_{1,}t_2\right) = C - PATH\left(t_{1,}t_2\right) - K * dir \qquad (3.11)$$

where dir is the number of directional changes from two terms t1 to t2, and C, K are constants whose values are based on experiments.

3.4. Update of Feature Weights

After obtaining the reduced functionality is achieved, a new weight is given to the terms depending on the optimized semantic similarity metric applied. Once the semantic measure is computed, a match between two terms is generated. If that distance is below a certain threshold, then the weight of this term is refreshed with a new weight value computed by:

$$F = \left(w_i, w_j\right)X(1 - dist_{ij}) \qquad (3.12)$$

In which w_i and wj are the weights (TF-IDF) of two terms i, j and $dist_{ij}$is the length of similarity between the two terms.

3.5. Enriched Firefly Optimization Algorithm for Feature Selection

All fireflies in the search area is performed with a random number in the [0,1] range. The of Each firefly’s position is computed with the eq.(3.13)

$$x = [v - 1]x\, Rand[0,1] + 1 \qquad (3.13)$$

Where v = top boundary [1,0] and l=bottom boundary [0,0].

The resulting sequence is transformed into a binary sequence by means of the relation in eq (3.14).

$$b_i = \begin{cases} 1, sigmoid\ (x) > v[0,1] \\ 0, otherwise \end{cases} \qquad (3.14)$$

Where x_i = location of each firefly, 1 = chance of one function being selected, and 0 = likelihood of a feature not being selected. Each Firefly initiated in the swarm has its own location according to the number generated from each Firefly [4]. In the proposed algorithm, the FF is determined in a way that minimizes the classification error rate on the validation dataset, as demonstrated by eq (3.15).

$$Error = 100 - A \qquad (3.15)$$

Calculation of every firefly’s attractiveness. To compute the degree of attraction β of every Firefly eq.(3.16).

$$\beta(r) = \beta_0 \times e^{-\gamma r2} \qquad (3.16)$$

where r = the distance between 2 fireflies , β_0 = the attraction of one firefly in the first case (r = 0).

$$r_{ij} = |x_i - x_j| \qquad (3.17)$$

where X = legitimate position values of each firefly previously calculated using equation of data gain ratio.

This distance between two fireflies is computed using the hammering distance method, in which every bit of firefly is deducted from the firefly. In this way, the distance is shown as the difference between the binary strings of the both fireflies [6]. This method enhances the ability of the EFA to work more effectively better with binary features than it does with continuous values.

In the swarm, every firefly is swallowed up by a brighter fireflies. In the algorithm, the best firefly position is updated by means of equation (3.18).

$$x_i = x_i + \beta x\left(x_j - x_i\right) + \alpha \times (Rand - 0.5) \qquad (3.18)$$

A two-dimensional stress state consists of three diferent stress components.The normal stresses σyy and σxx and the shear stress σxy = σyx. Randomness is reduced by different constant rate δ, where δ ∈ [0.95,0.97] so that at the final level of optimization, the value of α will be maximised, as in equation (3.19).

$$\alpha = \alpha \times \delta \qquad (3.19)$$

EFOA Algorithm

```
Step 1: The spam dataset is preprocessed and normalized.
Step 2: Extract the Email features and get synonyms set of a term.
Step 3: Merge the terms that have common synonyms and increase its weight.
Step 4: Reduce the feature using semantic based path reduction.
Step 5: Calculate the similarity distance between terms for each measure.
Step 6: Calculate the new weight and updated weighted feature.
Step 7: Selecting the best feature using Fireflies.
Step 8: Initialize all swarm of n Fireflies.
Step 9: Calculate the fitness and the light intensity of each firefly using.eq.(3.16).
Step 10: Generate spam optimal topologies using stress method.
Step 10: Update the position of best ($f_i$) using eq. (3.18)
Step 11: Get the swarm sorted and locate the best firefly.
```

4. Results and Discussion

A. Datasets

Spam dataset is intended to categorize e-mail as Spam or Spam-free.There are 4601 emails and 58 attributes in this dataset. Each instance within SPAM is made up of 58 attributes. Most attributes denote the frequency of a particular word or character in the email which matches the instance. The frst 48 attributes include the frequency of the attribute name within the e-mail. Attributes 49–54 are the number of characters _;‘, _(‘, _[‘, _! ‘, _$ ‘, and _# ‘. Attributes 55 to 57 define the average; longest and total duration of uppercase letters. Attribute 58 specifies the type of mail which is either "nonspam" or "spam".

Enron-Spam is a collection of e-mails consisting of six different datasets each containing ham messages from a single enron corpus user. In six datasets, most emails in Enron 1-3 are legitimate, whereas most emails in Enron 4-6 is spam. This dataset consists of 30,041 electronic messages [17].

B. Evaluation Metrics

Generalized and performing indices are necessary to assess the proposed algorithm in order to compare its results with those of other methods in this respect.

Accuracy

Accuracy is computed as the percentage of the dataset correctly caregorized by the algorithm. The percentage of total number of properly recognized e-mails defined by the following formula:

$$Accuracy = \frac{No\ of\ correctly\ classified\ non\ spam\ emails + No\ of\ correctly\ classified\ spam\ emails}{Total\ No\ of\ spam\ emais + Total\ No\ non\ spam\ emails} \tag{4.1}$$

Precision

It indicates the number of jurisdictions which are positively ranked and relevant. High precision demonstrates high pertinence for positive detection.

$$Precision = \frac{Correctly\ classified\ non\ spam\ emails}{Correctly\ classified\ non\ spam\ emails + Falsely\ classified\ non\ spam\ emails\ as\ spam} \tag{4.2}$$

Recall

Spam recall is defined as the probability of correctly classifying spam e-mails as spams, and the legitimate recall is defined as the probability of properly classifying correctly legitimate e-mails. The Recall formulas are listed below:

$$Recall = \frac{Correctly\ classified\ non\ spam\ emails}{Correctly\ classified\ non\ spam\ emails + Falsely\ classified\ spam\ emails\ as\ non\ spam} \tag{4.3}$$

False Positive Rate

The false positive rate defines the error in judgment ratio of legitimate messages as spam. The FPR formulas are listed below:

$$FPR = \frac{Falsely\ classified\ non\ spam\ emails\ as\ spam}{Falsely\ classified\ non\ spam\ emails\ as\ spam + correctly\ classified\ spam\ emails} \tag{4.4}$$

False Negative Rate

The false negative rate defines the spam mail misjudging ratio as legitimate. The FNR formulas are listed below:

$$FNR = \frac{Falsely\ classified\ spam\ emails\ as\ non\ spam}{Falsely\ classified\ spam\ emails\ as\ non\ spam + Correctly\ classified\ non\ spam\ emails} \tag{4.5}$$

Time

It measures the amount of time taken to filtering the spam emails from the database.

C. Results

The proposed method produces high accuracy when compared with previous method.The efficiency of SCA and EFOA is analysed and compared against existing methodologies, namely MLP [15] and SCA [14]. Table 1 shows, performance of filtering accuracy values of spam and non spam messages.

Table 1. Filtering of Spam and Non Spam Accuracy Measure for Spam Dataset

Methods	Spam Accuracy(%)	Non Spam Accuracy(%)	Over all Accuracy(%)
MLP	91.88	94.52	93.20
SCA	97.51	98.33	97.92
EFOA	**97.66**	**98.74**	**98.20**

The results shows that the EOFA method shows better filtering of spam and non spam accuracy values when comparing with other existing methods in spam dataset.

Table 2. Filtering of Spam and Non Spam Accuracy Measure for Enron Spam Dataset

Methods	Spam Accuracy(%)	Non Spam Accuracy(%)	Over all Accuracy(%)
MLP	91.12	93.52	92.32
SCA	92.39	93.97	93.18
EFOA	**95.94**	**97.62**	**96.78**

The results shows that the EOFA method shows better filtering of spam and non spam accuracy values when comparing with other existing methods within enron spam dataset.Table 3 shows, the performance assessment for the spam datasets.

Table 3. Performance Measure for Spam Dataset

Methods	Precision (%)	Recall (%)	False Positive Rate	False Negative Rate	Time (Sec)
MLP	93.20	91.47	0.184	0.211	12
SCA	98.64	96.52	0.132	0.166	10.4
EFOA	**98.92**	**97.03**	**0.058**	**0.098**	**7.4**

The EFOA method on spam dataset has better precision, Recall, False Positive Rate,False Negative Rate, time than EFOA method on enron spam dataset.Table 4 shows, performance evaluation of enron spam dataset.

Table 4. Performance Measure for Enron Spam Dataset

Methods	Precision (%)	Recall (%)	False Positive Rate	False Negative Rate	Time (Sec)
MLP	91.94	88.91	0.206	0.279	13.2
SCA	94.10	92.16	0.153	0.197	11.9
EFOA	**97.19**	**95.28**	**0.076**	**0.142**	**8.8**

This results shows the comparison between spam and enron spam datasets in terms of Precision,Recall,False Positive Rate, False Negative Rate, Time on different Methods. From this analysis, it is proved that the proposed EFOA takes less time than SCA for Email spam filtering on spam and enron spam datasets.

5. Conclusion

In this article, EFOA is proposed to effectively deal with the space and time complexity problem in the context of email spam filtering system. The optimized semantic WordNet is employed to clean the noise data and reduce the large number of text features extracted from the email. In optimized semantic WordNet, Semantic-based reduction and Features weights updated are used for better feature reduction and representation. After selecting best features are fed into ANN classfier technique and classify the spam and non-spam of email messages. The experiments are carried out in a spam dataset and the experimental results demonstrate that the proposed EFOA is better than the existing SCA method.

References

1. Mujtaba, G., Shuib, L., Raj, R.G., Majeed, N., & Al-Garadi, M.A. (2017). Email classification research trends: review and open issues. IEEE Access, 5, 9044-9064.
2. Idris, I., & Selamat, A. (2014). Improved email spam detection model with negative selection algorithm and particle swarm optimization. Applied Soft Computing, 22, 11-27.

3. Shuaib, M., Abdulhamid, S.M., & Adebayo, O.S. (2019). Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification.
4. Yang, X. S., & He, X. (2013). Firefly algorithm: recent advances and applications. International journal of swarm intelligence, 1(1), 36-50.
5. Kakade, A.G., Kharat, P.K., Gupta, A.K., & Batra, T. (2014). Spam filtering techniques and MapReduce with SVM: A study. In 2014 Asia-Pacific Conference on Computer Aided System Engineering (APCASE), 59-64.
6. Ahmed, H.A., Zolkipli, M.F., & Ahmad, M. (2019). A novel efficient substitution-box design based on firefly algorithm and discrete chaotic map. Neural Computing and Applications, 31(11), 7201-7210.
7. Li, W., & Meng, W. (2015). An empirical study on email classification using supervised machine learning in real environments. In 2015 IEEE International Conference on Communications (ICC), 7438-7443.
8. Mallik, R., & Sahoo, A.K. (2019). A novel approach to spam filtering using semantic based naive bayesian classifier in text analytics. In Emerging technologies in data mining and information security, 301-309.
9. Zhang, D., Yan, Z., Jiang, H., & Kim, T. (2014). A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. Information & Management, 51(7), 845-853.
10. Laorden, C., Ugarte-Pedrero, X., Santos, I., Sanz, B., Nieves, J., & Bringas, P.G. (2014). Study on the effectiveness of anomaly detection for spam filtering. Information Sciences, 277, 421-444.
11. Awad, M., & Foqaha, M. (2016). Email spam classification using hybrid approach of RBF neural network and particle swarm optimization. International Journal of Network Security and Its Applications (IJNSA), 8(4), 17-28, 2016.
12. Sharma, A.K., & Yadav, R. (2015). Spam mails filtering using different classifiers with feature selection and reduction technique. In 2015 Fifth International Conference on Communication Systems and Network Technologies, 1089-1093.
13. Hristea, F.T. (2013). Semantic WordNet-based feature selection. In The Naïve Bayes Model for Unsupervised Word Sense Disambiguation, Springer, Berlin, Heidelberg, 17-33.
14. Pashiri, R.T., Rostami, Y., & Mahrami, M. (2020). Spam detection through feature selection using artificial neural network and sine–cosine algorithm. Mathematical Sciences, 14(3), 193-199.
15. Abdulhamid, S.I.M., Shuaib, M., Osho, O., Ismaila, I., & Alhassan, J.K. (2018). Comparative Analysis of Classification Algorithms for Email Spam Detection. International Journal of Computer Network & Information Security, 10(1). 60-67.
16. Sharma, M., & Sharma, S. (2018). A Survey of Email Spam Filtering Methods. Control Theory and Informatics, 7, 14-21.
17. Sharma, P., & Bhardwaj, U. (2018). Machine learning based spam e-mail detection. International Journal of Intelligent Engineering and System, 11(3), 1-10.
18. Bassiouni. M., Ali. M., & El-Dahshan. E.A. (2018). Ham and Spam Email Classification Using Machine Learning Techniques. Journal of Applied Security Research, 13(8), 315-331.
19. Rekha, S.N. (2014). A review on different spam detection approaches. International Journal of Engineering Trends and Technology (IJETT), 11(6), 315-318.
20. McInnes, B.T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of biomedical informatics, 46(6), 1116-1124.
21. Bahgat, E.M., Rady, S., Gad, W., & Moawad, I.F. (2018). Efficient email classification approach based on semantic methods. Ain Shams Engineering Journal, 9(4), 3259-3269.