ISSN Online: 2821-1693

Transactions on Machine Intelligence

Journal Homepage: https://tmachineintelligence.ir/index.php/TMI

Spam Detection from Big Data based on Evolutionary Data Mining Systems

H. Ehsani Chimeh[1,*], M. Karami[2]

[1] Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran
[2] Department of Electrical Engineering, Shahid Beheshti University, Tehran, Iran

ARTICLE INFO

Article History:
Received 29 January 2018
Received in revised form
14 February 2018
Accepted 4 March 2018
Available online 11 March 2018

Keywords:
Spam Detection, Big Data, Genetic Algorithm, Self- Organizing Neural Network (SOM), Probabilistic Neural Network (PNN)

ABSTRACT

News releases and users' ability to discuss events, events, and writing personalities and environments are services that provide opportunities for new types of spam and spammers. For example, popular topics and topics that involve the most discussions can be an opportunity to create traffic, visits, and sources of income. When something happens, thousands of users write about it, send text and quickly become the subject of discussion. These topics are targeted by spammers, because their writings contain the common words used in popular discussions. Often there are links in spam that direct users to websites that are not related to the topic, and since these URLs are shortened, it's difficult for users to log in. This type of spams can reduce the value and efficiency of instantaneous search services, and users of these services refer to materials that do not contain links to the searcher, so a method for identifying spammers should be found. Methods available to deal with spammers can be included in three categories which contain detection-based approach, prevention-based approach, and degradation-based approach that this research uses is a detection approach. Hence, this research uses a smart method that initially enters large data into the program, then a feature extraction based on the genetic algorithm is performed. In the next step, the classification of data in order to detect spam is done using the combined method of self-organized mapping neural network and probabilistic neural network with the support vector machine core as a radial basis function.

1. INTRODUCTION

Today, online social networks and e-mail are one of the fastest and most economical ways to communicate. However, the increase in users of online social networks and email has led to an unprecedented increase in the number of spam in recent years. Spammers are always looking for new victims by sending unwanted messages to other users. These malicious users are just looking for their goals. These goals can be marketing, advertising with specific orientation or disturbing communications. Spamming has many forms, such as sending spam messages in mail or asking for unwanted friendships in messenger systems or online social networks [1].

* Corresponding Author: hrechime@gmail.com
Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran

http://dx.doi.org/10.47176/TMI.2018.1

Spammers on social networks try to increase the number of users who communicate with them for the rapid dissemination of their messages, which often contain propaganda and fraudulent content and analytically of low value, as well as because identifying their account by network or other users will alternate create a new account. These cases cause a lot of communication in the network graph and nodes that are not a proper mapping of the real world and will negatively affect the analysis of social networks based on the structure of the graph. Also, the content sent by these users also negatively affects message-based analysis methods [1]. The existence of spammers on social networks makes the structure of the network a BIAS and influences the analysis of social networks in various fields. The BIAS on social networks is an estimator, the difference between the mathematical hope of the estimator and the actual value of the estimated parameter. Therefore, providing an intelligent method that has the ability to detect spam is a necessity.

2. LITERATURE REVIEW

Spam traffic specifications vary from authorized traffic authorization specifications. Authorized letters are usually issued throughout the day, while spamming is uniform throughout the day [2]. Spammers keep their identities secret when they send spam, but their identities are detectable when hunting mail addresses from websites, and this is one way of identifying spammers on the Internet. Spamming continues to be economically viable as advertisers do not spend any time managing their mailing lists, and this makes it harder for the senders of the letter to be held responsible [3].

Several methods have been developed to deal with unwanted electronic mail, which can be used to refer to filtering or filtering software. The first filters looked superficially, just checking the existence or absence of a series of predefined words in the body of the message, or whether they were referring to the sender, whether it was a white list or a blacklist. The whitelist is known as addresses, and there is a tendency to receive mail from them, and the blacklist is also a collection of addresses that are unwilling to receive mail from them. These two methods have certainly not been based on learning methods, but later on the idea of the two methods was used in learning-based methods. On social networking sites, there are several approaches to dealing with spam. According to [4], these approaches can be categorized in three different categories, based on diagnosis based on prevention, based on rating degradation.

In [5], they also used a completely similar method to detect spammers on Twitter. Twitter's social network allows users to post content to the network, and people who follow the user to view the content and re-register if they wish to. In this way, an item in the network will be expanded and sent to other users. This makes perpetrators join the network for posting their content. In [6], the pseudoscientific data set on the Twitter website and Facebook is used to detect spam based on the genetic analysis algorithm. Also, the Jrip decision tree algorithm has also been used as a feature extraction and classification method.

In [7], the detection of spam from the blogging system from blog search results has been researched, which provides a new framework for monitoring user behavior in a blog commentary that is based on observation learning methods. In [8], the Spam Detection Tool is based on social cascading information and follows it to spammers, based on the Twitter data set. The proposed method is based on the theory of the situation. In [9], spam and spammers detect blogging systems with the benefits of social networking content that is based on graph structures. This system has the ability to detect spam and spammers at the same time. In another research, users' comments in a polling and scoring system that is spam has been analyzed [10]. The method used is a binomial regression that is presented on a user's view of a product.

The use of self-organizing neural network as a method for detecting spam is presented as a new approach in [11], which works hierarchically. This system is used in internet distribution networks or ISPs. [12] A spam detection method and abnormal operation in a network with a neural network approach. The use of the self-organized mapping neural network is considered as data training and feature extraction as well as the use of an optimized Particle Swarm Optimization (PSO) algorithm to classify spam types and spoof the network. In [13], spyware detection on the Twitter community has been conducted with a hybrid algorithm called SPD, whose results are meaningful.

3. PROPOSED METHOD

At first, data should be normal which this dataset used in this project is normal. Therefore, it is necessary to extract the features for the next operation. In order to extract the attribute, the genetic algorithm will be used. In order to deal with the high input characteristics, feature selection is used to reduce the size and identify the most relevant attributes that can cause the separation of different classes to be sufficient. The method of selecting a feature is a sensitive method, which is why inadequate features reduce the classification efficiency, while a larger set of features does not always result in identifying the results more accurately. The genetic algorithm uses a series of operators including the initial population to produce chromosomes and genes and then with the some operators such as crossover, mutations, and selections, in a given repeat interval based on fitness function, finds any spam is big data. In fact, this section has a text-mining mode based on the data mining model.

The genetic algorithm approach is such that M represents the number of educational data, the average data F_i and $l_i$of each data from the vector T_i. First, there is a number of data, each of which contains $N \times N$ dimension. Each data can be represented in a N-dimensional space whose relation is in the form of equation (1) and the averaging operation is carried out in the form of equation (2).

$$A = N \times N \times M \tag{1}$$

$$F_i = \frac{1}{M}\sum_{t=1}^{m} Tt \tag{2}$$

Finding the standard deviation is important issue in the spam detection system which considered to be big data with a genetic algorithm calculated through equation (3) and the covariance matrix is also calculated from equation (4).

$$Variance = \frac{1}{M}\sum_{t=1}^{m} Tt \tag{3}$$

$$Cov = AA^T \tag{4}$$

Where $A = \lfloor Variance_1 . Variance_2 . \ldots . Variance_n \rfloor$ and $Cov = N^2 * N^2$ are a matrix, because $A = N^2 * M$ is a matrix. So Cov is a huge amount. We now obtain special values from Cov using equation (5).

$$U_i = AV_i \tag{5}$$

The final stage is the selection of the special vector. A set of features of the inherent state function, which is represented as $N(N = 213)$ in a d-dimensional space $\{x_1 . x_2 . \ldots . x_N\}$, which is $x_i = R^d$ and belongs to $C(C = 7)$ The class of $\{L_i i = 1.2. \ldots . C\}$ is available. The goal of the genetic algorithm is to search linear transitions for mapping the original d-dimensional space to f-dimensional, which is $f < d$. The new feature vector is located at $y_i = R^f$. The scattered matrix is given in the class as the total dispersion or covariance matrix calculated as equation (6) and (7).

$$S_T = \sum_{k=1}^{N} (x_k - \mu)(x_k - \mu)^T \tag{6}$$

$$W_{GA} = arg\ max[W^T S_T W] = [w_1 w_2 \ldots w_f] \tag{7}$$

According to (6) and (7), μ is the average of all the samples and $\{w_i | i = 1.2. \ldots . f\}$ is a set of special vectors f-dimensional of S_T, which is associated with the largest eigenvalue f. The view of the samples in the new space is $y = W^T x$, where $W_{GA} \in R^{fxd}(170xd)$. As the starting point in the production of primary populations of chromosomes, the educational data are read in dimension $N \times N$ and converted to $N^2 \times 1$ dimension. A training set in dimension $N^2 \times M$ is also made, where M is the number of data samples. The mean of the dataset is calculated by the equation (8).

$$\psi = \frac{1}{M}\sum_{i=1}^{M} \Gamma_i \tag{8}$$

The class of equation (8), ψ is the mean of data, M is the number of data and Γ_i is the data vector. The combination of the corresponding chromosomes is retained with the highest specific amount. This combination defines the space of the offender. Special space is created by plotting data into a predator space that creates new combinations. Therefore, weight vectors are calculated based on the mutation. Then weighing the vector and weighing the parser weight in the database are compared. The average data is calculated and then subtracted from any data in the training set. Matrix A is constructed using the results of subtraction operations. The difference between each spammers and the mean of data is calculated as the equation (9) as the selection operation with a random structure.

$$\phi_i = \Gamma_i - \Psi . \qquad i = 1.2. \ldots . M \tag{9}$$

According to equation (9), ϕ_i is the difference between the predator and the average data. The matrix obtained by the subtraction operation, which is the same as the matrix A, is multiplied in its transposition, and finally the covariance matrix C is formed, whose relation is in the form of equation (10).

$$C = A^T A \tag{10}$$

According to equation (10), this A forms the difference between vectors, for example, we can mention $A = [\phi_1 . \phi_2 . \phi_3 . \ldots \phi_M]$. Dimensions of the matrix C are $N \times N$. The number of data samples or M is used to form a C matrix. In practice, we can say that the matrix C is $N \times M$. On the other hand, when A is equal to M, only M of N is the number of special numbers equal to the non-zero value. Then the covariance matrix values are calculated. The selected vector of special vectors is multiplied by matrix A to reduce the component space. Special vectors of smaller values correspond to smaller variations in the covariance matrix. Other features of the data are maintained. After dimming, selecting features and extracting features, it is necessary to train, classify and classify these data. Hence, the neural network of self-organizing mapping is based on a probabilistic neural network, Probabilistic neural networks have been used in many studies [14].

The self-organized neural network, the topological structure among the different input units assumes that this problem is not seen in other neural networks. In these networks, n input signals are assumed to be in the m cluster and the clusters are arranged in a one-dimensional or two-dimensional arrangement of regular. The vector of weight for each cluster is a sample vector of input patterns linked to that cluster. A neural network is a kind of neural network of radial base function whose spread function is considered in probabilities. The potential neural network consists of four layers: the input layer, the pattern layer, the consensus layer, and the output layer.

In equation (11), σ is the smoothing parameter, X_{ij} is the neuron vector, and d represents the vector dimension of the pattern. The consensus layer in order to obtain an estimate of the probability density of each mode and its output is proportional to the estimation of different units of probability density based on the core. The sum of the maximal likelihood of the X pattern being classified in C_i is calculated by summarizing and averaging the output.

$$\emptyset_{ij}(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \; exp\left[-\frac{(x-x_{ij})^T (x-x_{ij})}{2\sigma^2}\right] \tag{11}$$

In equation (12), N_i represents the total number of samples in the class C_i. If there is a probability of forecasting for each class, and the losses associated with making an incorrect decision for each class, the decision-making unit classifies the x pattern according to the decision-making law of the business on the basis of the output of all the neurons in the aggregate application.

$$p_i(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \; \frac{1}{N_i} \sum_{j=1}^{N_i} exp \; \left[-\frac{(x-x_{ij})^T (x-x_{ij})}{2\sigma^2}\right] \tag{12}$$

The output layer selects the maximum probability density of the neurons as the total output of the system in the probability density estimation. Output layer neurons are a kind of competitive neurons that each neuron belongs to and is a type of data that is part of the classification. Two neural networks combining a self-sustained and probabilistic mapping neural network require a common core. Hence, they are used as the core of the backup vector machine. The core type is also a radial base function for combining two neural networks.

4. SIMULATION AND RESULTS

One of the most important issues in this research is the data collection, which will be Twitter on Twitter in November 2017. This dataset has 1000 user data that has 4 attributes. It is planned to use up to 70% of data as training and 30% of data as a test set. Using this proposed algorithm, select 11 of the 17 features in the dataset, the entire set of which is:

1. Number of characters per post (minimum)
2. Number of characters per post (maximum)
3. Number of characters per post (middle)
4.Number of characters per post (average)
5. Number of tags per post (maximum)
6. Number of web addresses per post (minimum)
7. Number of web addresses per post (max)
8. Number of characters per post (average)
9. Number of web addresses per post (average)
10. Number of times the posts have been re-released (average)
11. Number of posts
12. Number of follower followers
13. Time interval between posting (average)
14. Number of posts per day (middle)
15. Number of posts per day (minimum)
16. Number of posts per day (average)
17. Account lifetime

After the extraction operation, the self-organized mapping neural network is first taught with educational data, and then the evaluation data is presented to improve the classification to the probabilistic neural network and the results are extracted. During the initial classification with the self-organized mapping neural network, the results of the evaluation are evaluated based on the extraction of the characteristics based on the genetic algorithm. In the results, a good percentage of reputable users was identified, but this method has had two main problems:

The accuracy of the identification of ordinary users is 80.56%, but this means a high error of about 19.46% in the identification of ordinary users, which means 19.46% of ordinary users are classified as a perpetrator. This is not suitable for the system, because the misleading classification of ordinary users can reduce the users of the social network.

In the proposed method, the feature selection has been used and 17 features that have more suitable information for the classification of users have been extracted, but this has a major disadvantage for the system of the system presented in this study, because in This particular issue should be used by all the features to classify users.

Users have several different feature categories, some of which are user-friendly, for example, the user can pay more attention to the number of web page addresses in his posts, in other words, the user in the features he controls It can try to make its behavior more similar to ordinary users and confuse the detection system, so it's best to use all the features to identify users and not to choose features for users.

In the current issue, one of the biggest challenges is the dual behavior of users, users who in some cases behave as patrons, and behave like normal users in the rest of the world. You should focus on identifying these users. Hence, it is important to use a method that can optimize classed classes. On the other hand, the purpose of using the self-organizing mapping neural network, which is classified as a classifier, must be in some way combined with a neural network of self-organizing mappings based on a probabilistic neural network. Finally, by examining the data and results of the methodology and the exact examination of the data, it is most mistaken to classify users who have a dual behavior. These users behave normally in most cases and behave like regular users, but often send messages, on the other hand, they are users who, although not spammers, are behaving like perpetrators. These two groups of users caused a very high error in the previous methods. In order to determine the behavior of these users, it is necessary to identify the users in the training data of the users and model their behavior separately from other users. In this research case, there are two models for users, a model for identifying users with dual behavior and a model

for users whose behavior is not dual, in other words, there is a model for identifying users trying to deceive spyware identification systems And a model for identifying users who behave independently of the system of identifying spammers. In assessing users, you first have to determine whether the user has dual behavior or not, if the user has dual behavior, the corresponding model is used for identification, otherwise it will be used by another model.

Many experiments have been done to find the number of neurons in the secret layer in the proposed self-organizing mapping neural network. The hidden layer is a layer whose existence is necessary because the patterns are divided into several classes. The self-organizing neural network is trained by Levenberg Marquardt. Levenberg Marquardt algorithm for multi-layer networks, especially the self-organized mapping neural network, is an extension of the LMS algorithm, and both have the same performance index, which is the mean squared error. This algorithm reduces the mean squares of errors between the desired output and the actual output, which is a function of the stimulus used, of the Levenberg Marquardt type, known as the TRAINLM. Like Gauss-Newton's method, Levenberg Marquardt algorithm is designed to approach quadratic education without computing the Hessian matrix.

The network to be used for training uses two layers, both of which use the Lewenberg Marcard stimulator function with 2 neurons. The first layer uses the sigmoid tangent actuator function and the second layer uses a linear actuator function. The shape of the self-organizing neural network is shown by combining the probabilistic neural network in Fig. 1.

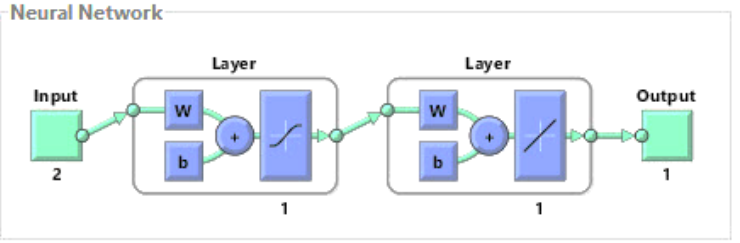

Fig. 1. Structure of the compound neural network

The number of neural network cycles is considered to be 1000 rounds. The network mutation rate is also 0.001. The weight of each layer is also 1. In order to train 70% of input data as training and 30% as a test, the neural network performance in Fig. 2 as well as different learning modes in Fig. 33 and regression are also shown in Fig. 44. Is.

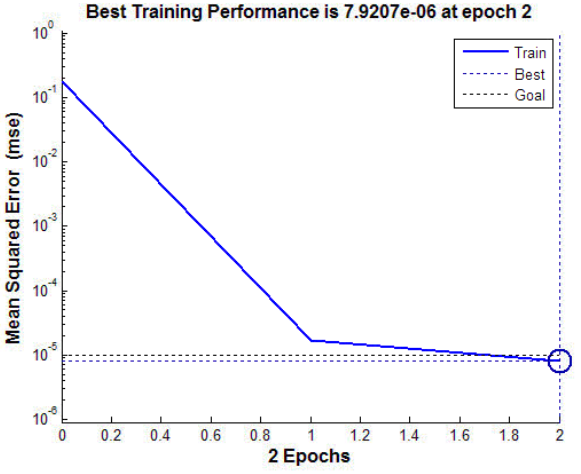

Fig. 2. Performance of combined neural networks

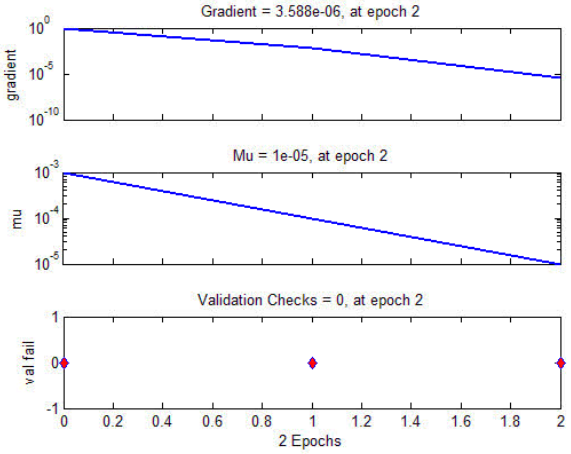

Fig. 3. Training states of combined neural networks

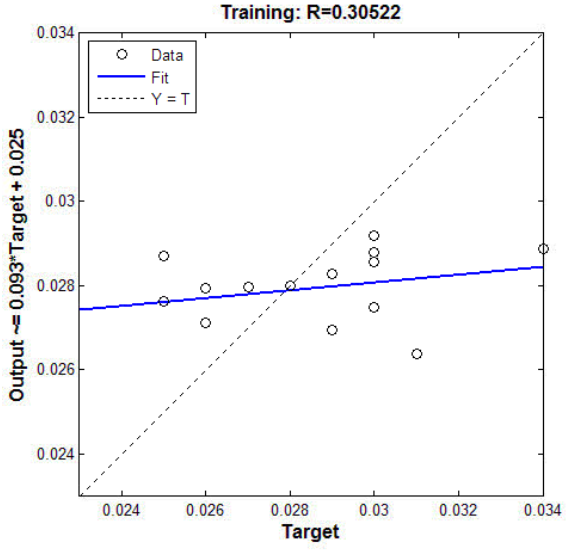

Fig. 4. Regression of combined neural networks

Now that simulation has been performed, it is necessary to measure the obtained rate using several evaluation criteria. When the simulation finishes, results are obtained. The resultant bias classification is 0.1309. A total of 99 data were identified as the main input, of which 70% were used as training and 30% were used as a test, 50 offspring and 49 normal users were identified. We've named the coworkers with the correct results and normal users with the wrong results. It should be noted that the classification error based on the proposed algorithm is 0.0695%, which is negligible. Now that information is derived from these methods, these methods apply to the results obtained by the classification to identify the offender. The results of this work are shown in Table 1.

Table 1. Evaluation results of proposed method

Sensitivity (%)	Accuracy (%)	MSE on Training	MSE on Test
89.0909%	90.91%	0.1648	**0.1563**

From the results, it is clear that the proposed method has very good results at the time of the discovery of spammers, and also the results of the evaluation show the robustness of the proposed method. The proposed method is compared with two other similar researches in terms of functional accuracy in terms of percentage for detecting spammers and, in general, the accuracy of the program in the classification, the result of which is shown in Table 2.

Table 2. Comparison of the proposed method with other related researches

References	Accuracy
Benevenuto, Fabricio et al., 2010 [5]	83.01 %
Ahmed, Faraz, and Abulaish, Muhammad, 2013 [6]	89.06 %
Isa, Inuwa-Dutse, et al., 2018 [19]	89.70 %
Proposed Method	90.91 %

Of course, it cannot be accurately stated that the comparisons are so correct, since the number of data used in the training phase, as well as the test, the data set and for which social network they are, is important. But overall, based on accuracy, a comparison has been made between the three similar studies, which shows that the proposed method has the ability to detect spammers with a precision of 90.91%.

5. CONCLUSION

Today, the Internet has been approved as a useful tool for user-friendly use based on user requirements. One of the applications that uses the Internet is social networks that can be used with a wide range of goals. Dating, sending textual and multimedia data quickly to people around the world, as well as online conversations with people, including social networking features. One of the social networking sites that are being approved by the users is based on the features that it offers, the Twitter community. One of the major challenges and challenges in these networks is the existence of intrusive writings that they write and send to some, which, in addition to interruptions, leads to increased network traffic and reduced bandwidth. Impairs access. These spam-known writings, as their name suggests, are created as weavers throughout the network, starting with a communication channel that is a systemic problem for social networks, and is essentially a programmer discovering it. It sends data. People who produce such data are known as spammers. Identifying and detecting spam and spammers is very important in social networks in order to benefit from the use of the network for users. In this research, we use the combination method of genetic algorithm to extract the characteristic and further classification with the characteristics of the result based on the combination of self-organizing mapping neural network and probabilistic neural network with support core machine with radial base function. The results show the performance and high accuracy of the proposed method compared to other methods.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Wu, T., Wen, S., Xiang, Y., & Zhou, W. (2018). Twitter spam detection: Survey of new approaches and comparative study. *Computers & Security, 76*, 265–284. https://doi.org/10.1016/j.cose.2017.11.013

[2] Fei, G., Li, H., & Liu, B. (2017). Opinion spam detection in social networks. In *Sentiment analysis in social networks* (pp. 141–156). https://doi.org/10.1016/B978-0-12-804412-4.00009-7

[3] Eshraqi, N., Jalali, M., & Moattar, M. H. (2015). Spam detection in social networks: A review. In *2015 International Congress on Technology, Communication and Knowledge (ICTCK)*. https://doi.org/10.1109/ICTCK.2015.7582661

[4] Chakraborty, M., Pal, S., Pramanik, R., & Ravindranath Chowdary, C. (2016). Recent developments in social spam detection and combating techniques: A survey. *Information Processing & Management, 52*(6), 1053–1073. https://doi.org/10.1016/j.ipm.2016.04.009

[5] Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on Twitter.

Proceedings of the 6th Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS).

[6] Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in online social networks. *Computer Communications, 36*(10–11), 1120–1129. https://doi.org/10.1016/j.comcom.2013.04.004

[7] Zhu, L., Sun, A., & Choi, B. (2011). Detecting spam blogs from blog search results. *Information Processing & Management, 47*(2), 246–262. https://doi.org/10.1016/j.ipm.2010.03.006

[8] Jeong, S., Noh, G., Oh, H., & Kim, C.-K. (2016). Follow spam detection based on cascaded social information. *Information Sciences, 369*, 481–499. https://doi.org/10.1016/j.ins.2016.07.033

[9] Wu, F., Shu, J., Huang, Y., & Yuan, Z. (2016). Co-detecting social spammers and spam messages in microblogging via exploiting social contexts. *Neurocomputing, 201*, 51–65. https://doi.org/10.1016/j.neucom.2016.03.036

[10] Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2015). Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications, 42*(22), 8650–8657. https://doi.org/10.1016/j.eswa.2015.07.019

[11] Palomo, E. J., Domínguez, E., Luque, R. M., & Muñoz, J. (2009). Spam detection based on a hierarchical self-organizing map. In *Emerging intelligent computing technology and applications. With aspects of artificial intelligence* (pp. 30–37). *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-642-04020-7_4

[12] Shahreza, M. L., Moazzami, D., Moshiri, B., & Delavar, M. R. (2011). Anomaly detection using a self-organizing map and particle swarm optimization. *Scientia Iranica, 18*(6), 1460–1468. https://doi.org/10.1016/j.scient.2011.08.025

[13] Inuwa-Dutse, I., Liptrott, M., & Korkontzelos, I. (2018). Detection of spam-posting accounts on Twitter. *Neurocomputing, 315*, 496–511. https://doi.org/10.1016/j.neucom.2018.07.044

[14] Haddadnia, J., Seryasat, O. R., & Rabiee, H. (2013). Thyroid diseases diagnosis using probabilistic neural network and principal component analysis. *Journal of Basic and Applied Science Research, 3*(2), 593–598.