International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

A New Supervised Epidemic Model for Intelligent Viral Content Classification

Abdulkerim Senoglu[1], Uraz Yavanoglu[1], Suat Ozdemir[1]

Accepted 3rd September 2016

Abstract: In this study, we propose an information diffusion model, which is based on neural networks, artificial intelligence and supervised epidemic approach. We collected epidemically diffused data from Twitter with supervision to create a ranking system that forms the base of our diffusion model. The collected data is also used to train the proposed model. The outputs of the proposed model are shown to be useful for the provenance problem and the diffusion prediction systems in both physical and social networks. Knowing the viral content beforehand can be used in advertisement, industry, politics or any other end user that wants to reach a large number of people. Our performance analysis show that the proposed model can achieve over 90% training success rate and 78% test success rate of classifying viral content which is better than some of the existing models.

Keywords: *Artificial Neural Networks (ANN), Epidemic approach, Supervised learning, Information diffusion*

1. Introduction

Recently, mankind evolves with media, technology, and science to become more human than ever. We need to understand the world by using communication technologies. We are able reach almost any information in a matter of seconds. In this information age, other people's opinions reshape our perspective and decisions on certain matters about politics, health, religion and countless other elements. This is more likely the information provenance problem to understand data source [1]. In this problem, dissemination of information needs social pathways and number of peers to access by other network users [2]. This problem is proposed by Sola Pool and Kochen's but become well known with famous Milgram experiment in 1967 [3]. Milgram tried to understand average pathways and how information passing through random individuals. In his theory, six degrees of separation is enough to connect any two people [3]. The results basically suggested that small world networks can be characterized by short path lengths [3]. Information flow is mainly sharing believes and opinions in the same social circle [4]. This type of interaction creates a situation that is called homophily [5]. Influence and homophily are similar but distinctly affected states in social networking.

This study is related to information spread probability of individuals in social circles. One way to examine this kind of diffusion is epidemic modelling [7] which is based on diffusion like an epidemic spread of infectious diseases [6]. The epidemic modelling can be the key advantage to develop new information diffusion strategies to understand widespread of ideas, idioms and political views among geo-spatial social circles. In this study, we propose an information diffusion model that is based on neural networks, artificial intelligence and epidemic approach. To the best of our knowledge, this is the first study that employs neural network and epidemic model jointly to detect contextual information spread. We gather data from Twitter to create a ranking system that is the base of our diffusion model. The collected data is also used to train the proposed model. By using the proposed model we aim to classify viral content. The parameters of the are mainly adapted from the literature. However, two of these parameters, namely *verified account* and *challenge* are proposed are proposed in this paper. We achieve over 90% success for viral content classification rate and 0.0386 classification error rate which is better than some of the existing models. The rest of the paper is organized as follows; Section II presents related work. Section III explains the proposed model in detail. Finally, Section IV depicts the obtained results, conclusion and discussion.

[1] *Gazi University, Ankara, TURKEY*
** Corresponding Author: Email: abdulkerimsenoglu@gazi.edu.tr*
Note: This paper has been presented at the 3rd International Conference on Advanced Technology & Sciences (ICAT'16) held in Konya (Turkey), September 01-03, 2016.

2. Related Work

There is a significant amount of work on information spread in social networks. In this section we summarize the most notable works in area.

Jin, Fang et. al. studied epidemiological diffusion of news and rumors on Twitter. They used epidemiological models to characterize information cascades in Twitter that results from news and rumors. Epidemic model they used is SEIZ (susceptible, exposed, infected, skeptic) enhanced model to characterize events [7].

Yang, Jaewon, and Jure Leskovec developed a Linear Influence Model that predicts global influence of a node in the network on the rate of diffusion. In their work, every node has an influence function that quantifies how many infections can be related to the influence of that node over time. They demonstrated that Linear Influence Model find influenced nodes accurately [8].

Leskovec et. al. presented an analysis of a recommendation network. They monitored the diffusion of recommendations and cascade sizes and explained it with a stochastic model. They also considered the differences of user behavior within communities defined by recommendation network. They explained how the

recommendation network grows over time from the viewpoint of senders and receivers [9].

Bakshy, Eytan, et. al. considered the aspect of social networks in information diffusion. They inspected the role of social media in a field experiment that randomizes exposure to signals about friends' information. They found out that those who are exposed more likely to spread the information [10].

Aral, Sinan, and Dylan Walker studied identification of social influence in social networks. They proposed a method to identify influence and susceptibility in social networks. Their study revealed general influenced user groups on Facebook accounts. According to their work, younger users are more susceptible to be influenced, men are more influential than women [11].

Guille et. al. briefly studied the temporal dynamics of information spread in online social networks. They used machine learning based techniques to determine these dynamics. The results were represented on a real dataset that is extracted from Twitter and showed effectiveness of the proposed approach [12].

Bayzid et. al. proposed a new prediction model based on third world countries' problems. The authors constructed intelligent models with ANN methods to select suitable branch setup for microcredit organization and HIV risk determination of a locality in the third world countries. The results demonstrated that the ANN model predicts branch setup with 90% accuracy according to the feature vectors selected. The authors found out that feature vectors determining HIV/AIDS risk of a locality like microcredit problem. The authors claimed that this model might solve many local problems [13].

Thelwall et. al. constructed an intelligent model for sentiment analysis from MySpace members' comments with the help of data mining. The data mining approach consists of opinion mining or sentiment analysis phases. The methodology is to have opinion mining approach to gender differences in the expression of emotion applied for MySpace sites members' comments. The sample dataset contains containing 819 public comments or from 387 comments of U.S MySpace members. The experiment results showed that two thirds of the comments included positive emotion. In addition, females have more positive comments than males but there is no distinction for negative comments [14].

Junzhou Zhao et. al. studied the evolution of social information networks and how they are related to information spread. Study reveals a great portion of users has the ability to spread the information and people in a follower network tend to shorten the path of information flow [15].

Ghenname et. al. briefly studied the mass of data that social structures generate over time. They tried to take advantage of hashtag which is a labeling system to find relevant information in a mass of data. Their main focus is the labeling of hashtags for personalized recommendation on e-learning systems. [16].

Li et. al. studied recommendation systems and found out that even Recency(R), Frequency(F), and Monetary(M) (RFM) method and recommendation system combined they have difficulty increasing accuracy. They considered that taking product-purchasing timing in consideration could improve the accuracy and designed a system that considers purchase periodicity [17].

Shani et. al. combined two of the most used recommender system approaches; Collaborative Filtering and Content Based Filtering, to obtain better results from recommendations. The proposed algorithm is planned to work with different kinds of media type such as; audio, video, print etc. [18].

Diao et. al. considered one of the influential factors of recommending movies in a movie recommendation system is sentiment of the user whether it is negative or positive. They proposed a probabilistic model based on collaborative filtering and topic modeling. They evaluated that knowing the sentiment improves understanding the mechanism behind the developing rating systems [19].

Kong et. al., researched recommending appropriate level of cloud service to users with optimal compensation of their needs., They proposed a trust based recommendation system which takes direct trust and recommendation trust into consideration and unite them into trust value of their work [20].

3. Proposed Method

3.1. Artificial Intelligence

Artificial Intelligence (AI) is a field of scientific research to increase computing power, to develop productive algorithms and well organized knowledge. AI is applied for solving complicated problems that cannot be solved without combining intelligence, discovering the hidden patterns from data and developing intelligent machines [21].

AI has numerous applications on knowledge representation, information retrieval, speech recognition, understanding natural language, computer vision, bioinformatics, expert systems, robotics, game playing, and cyber defense with the help of various algorithms like artificial neural networks, genetic algorithms, artificial immune systems, particle-swarm intelligence, stochastic algorithms and fuzzy logic [22, 23].

Artificial Neural Network (ANN), which is a technique of AI, is defined as a mathematical model to imitate the human brain reasoning by examples [22]. Despite the fact that ANN have high accuracy even though noisy data by way of parallel computing, it has long training time and complex structure. ANN collects knowledge by detecting the relations among data and learns through its architecture and training algorithm.

Multilayered perceptron (MLP) network is the simplest ANN and for this reason commonly used for neural network architecture. MLP has three layers. Input layer, hidden layers and output layer respectively. The hidden layers' neurons are connected with all neurons in previous and next layers, and their connections are properly weighted [23]. The number of neurons in the input and output layers rely on the application, whereas neurons in the hidden layers are usually decided by trials [24].

Levenberg-Marquardt (LM) is a training algorithm to adjust neuron's weights by attempting to minimize the sum of squared error between the desired and actual values of the output neurons. LM is one of the most used learning algorithms because LM associates the Gauss-Newton technique with the steepest-descent method, besides LM is more robust than them and avoids many of their limitations [25].

In order to express this method, assume that w is parameter vector, is objective error function, is Jacobean that evaluated of f, error term (w) λ is Marquardt parameter and I is unit or definition matrix. The aim of the LM find the w when is minimum. Algorithm summarized as follows in Eqs. (1)-(4) [26];

$$E(w_k) = \sum_{i=1}^{m} e_i^2\,(w) = \|f(w)\|^2 \quad (1)$$

$$e_i^2(w) = (y_i - yd_i)^2 \quad (2)$$

$$\delta w_k = -\frac{J_k^T J_k\, f(w_k)}{J_k^T J_k + \lambda I} \quad (3)$$

$$w_{k+1} = w_k + \delta w_k \quad (4)$$

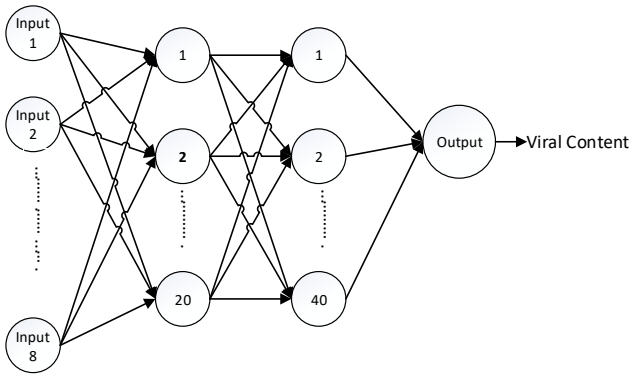

Fig 1. Proposed ANN model

3.2. Proposed Model

Our proposed model consists of two parts. First part is data collection which we used supervised learning to collect viral and non-viral data and label the parameters according to their value. Second part is intelligent epidemic model which classifies retrieved content into viral or non-viral classes with an ANN model.

Towards achieving the proposed model, a system is developed. The developed system is named as Intelligent Viral Content Classifier (I-VCC). Block diagram of ANN model which finds out best model for our study is shown in Fig.1. Steps of establishing the ANN model are briefly given below.

1) Data Collection: In this step, data is collected from various Twitter accounts in different time periods.

2) Pre-Processing: Raw data must be processed before labeling to detect related parameters and proper adjustments should be made.

3) Labeling: In this step, the parameters are labeled properly with their associated value and the data is prepared for the test and training phase.

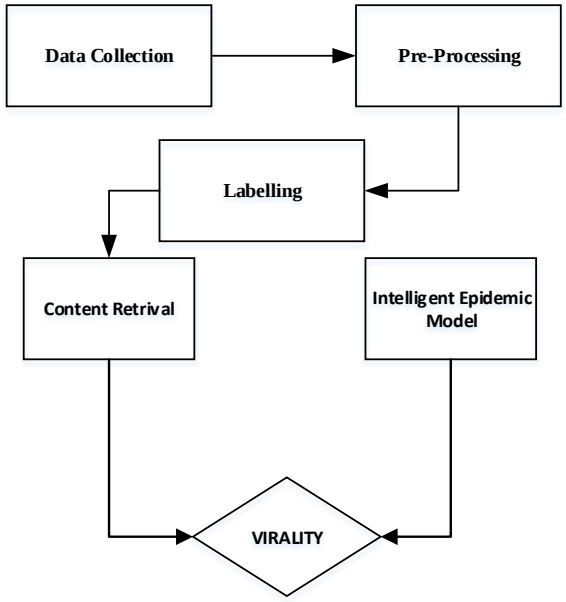

Fig. 2. Diagram of proposed I-VCC model with data collection

Steps of I-VCC are briefly given below.

1) Content Retrieval: In this step, the content is retrieved from web. Although we studied Twitter, other social sites are also applicable given the right parameters.

2) Intelligent Epidemic Model: The application asks for parameter values and the supervisor enters the proper values.

3) Virality: System predicts the given content is epidemically diffuses or not.

3.2.1. Data Collection

Tweets from various users and various dates and time used. Raw data collected from Twitter web site with the help of search, trending and user functions. Dataset consists of 300 tweets/users data. We collected tweets under supervision to determine the diffusion to be epidemically or not.

3.2.2. Data Pre-Processing

We processed the raw data in according to work with ANN. We extracted user and tweet information which we defined as ranking parameters. In our work, we used total of eight fields for ANN inputs, six parameters are inspired from literature such as user followers count [15], if trending (or contains hashtag) [16], timing [17], media type [18], content sentiment [19], if verified account [20] and two parameters are from our observation that; if contains challenge and if requests retweet. Parameters are explained with their associated table below.

3.2.3. Data Labelling

We labeled our parameters with their associated value. Values are differentiated in terms of parameter type. While binary parameters takes two arguments, others take more argument given their situation.

Table 1. Number of Followers

Follower Count	Value
1.000<	0
5.000>1.000	1
50.000>5.000	2
100.000>50.000	3
1.000.000>100.000	4
>1.000.000	5

Follower quantity is one of the most influential parameters. Although it is not enough by itself due to “follow me I follow you back” accounts, it makes a big difference when it comes to spreading information. We separated followers to 6 influence groups as shown in Table 1 [15].

Table 2. Time of Shared Content

Timing	Value
Other	1
10:00 - 14:00	2
16:00 - 18:00	3
23:00 - 01:00	4
19:00 - 23:00	5

Information sharing happens among social actors and it requires people who interacts with the information. Because of this reason timing parameter is added and separated 5 different time spans and given value according to Table 2 [17].

Table 3. Media Type

Media Type	Value
Text	1
Link	2
Gif	3
Video	4
Image	5

Media type is another huge factor for information sharing and it is also influential to spreading. Different media types have different social interaction and given values according to their influence as shown in Table 3 [18].

Table 4. Hashtag

Hashtag	Value
True	1
False	0

Hashtag is a labeling method and definitely increases one's shared content. Binary values given shown in Table 4 [16].

Table 5. Verified Account

Verified Account	Value
True	1
False	0

Verified account means whose information is true in their profile. Because of trust effect, verified accounts have more effect on information diffusion and the values deemed appropriate shown in Table 5 [20].

Table 6. Challenge

Challenge	Value
True	1
False	0

Challenge is a parameter that does not seen often but it has great effect to information diffusion. It means that someone challenges to beat a share count. We proposed this parameter because under right circumstances it can be only reason for a huge epidemic diffusion of information. Values for challenge parameter is shown in Table 6.

Table 7. Retweet Begging

Retweet Begging	Value
True	1
False	0

Retweet begging is another rare seen parameter but it also has a big effect on information sharing. People who lost their animal, patients searching for blood donor and even someone who wants attention and spreads fake news are retweet beggars and seen time to time in social media. We proposed this parameter because of increasing usage in social media sites from marketing to real needs. Parameter values are shown in Table 7.

Sentiment of a content or more clearly, positive or negative emotions have also effects on people to their sharing attitude. Positive emotions are almost always effects sharing positively and negative emotions have huge effect when people are depressed, feeling down because of a big event such as death of a national hero or a celebrity, natural disasters, terror attacks etc. We labeled sentiment effect into three parts as shown in Table 8 [19].

Table 8. Sentiment

Sentiment	Value
Positive	2
Negative	1
Notr	0

3.2.4. Data Sampling and Normalization

This process called as data sampling and normalization step to start analysis. For this purpose, we labeled data for ANN classifier to obtain more successful analysis result. We also labeled categorical data to numerical ones.

User followers count changes 0 to 5, if trending changes 0 to 1, timing changes 0 to 5, media type changes 0 to 5, if positive changes 0 to 1, if verified changes 0 to 1, if contains challenge changes 0 to 1, if contains negative content 0 to 1 and if requests retweet changes 0 to 1.

4. Results and Discussion

We collected 150 viral and 150 non-viral, total of 300 tweets from Twitter to test our proposed model. We performed 10 fold-cross validation to test our results. 10 fold-cross validation breaks data into 10 parts with n/10 size where n is total data size, then trains on 9 datasets and test on 1. This repeats for 10 times until all the dataset used. Experimental results show that ANN model has achieved 90% success rate for detecting epidemic outbreaks on given social media content.

4.1. Training

Training step consists of selecting the best transfer function, train model and epoch number from their different combination of different functions, models and numbers. 300 samples for training step and 30 samples for testing step are used. The most used transfer functions known as logarithmic sigmoid (L), hyperbolic tangent sigmoid (T) and linear function (P) are used as transfer functions and the most used train models known as Levenberg-Marquardt (LM), gradient descent (GD), Gradient descent with adaptive learning rate back propagation (GDA), Gradient descent with momentum and adaptive learning rate back propagation (GDX), Gradient descent with momentum back propagation (GDM) were selected as training model with different sizes of neuron count. Experimental parameters with minimum error for proposed I-VCC and ANN structure are shown in Table 9.

Table 9. Performances of Ann Models

ID	No of Neurons	Transfer Functions	Train Model	Epoch	Min Error
1	20,40,1	L,T,P	LM	70	0,0386
2	35,50,1	T,L,P	LM	67	0,0386
3	50,30,1	T,L,P	GD	82	0,0731
4	20,40,1	L,T,P	GD	87	0,0616
5	50,30,1	T,L,P	GDA	67	0,0442
6	20,40,1	L,T,P	GDA	64	0,0486
7	50,30,1	T,L,P	GDM	46	0,0729
8	20,40,1	L,T,P	GDM	49	0,0737
9	50,30,1	T,L,P	GDX	77	0,0398
10	20,40,1	L,T,P	GDX	75	0,0386

The best ANN structure with appropriate model and functions have 8 inputs, 20, 40 neurons and 1 output. LM model with T,L,P transfer functions gives the training performance shown in Fig.3.

Fig. 3. Estimated error rate

Our train results gives gradient value of 7.2935e-19 at 67 epoch as shown in Fig. 4.

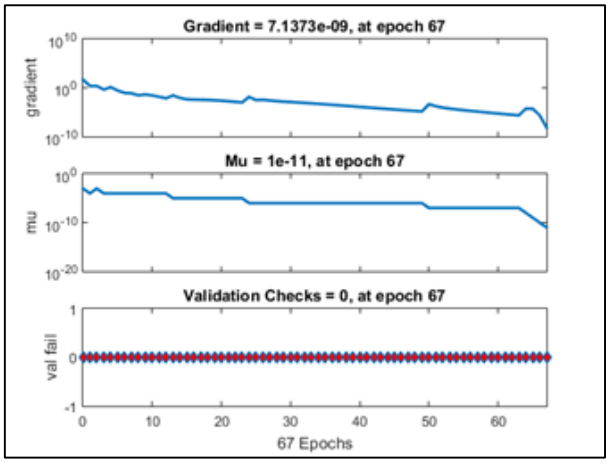

Fig. 4. Gradient, Mu and validation check graphs

4.2. I-VCC Tool

We developed an I-VCC tool for the proposed intelligent viral content classifier model. The tool takes arguments for the content parameters and classifies them as viral or non-viral as suggested in our model. A simple interface of the tool is shown in fig. 5.

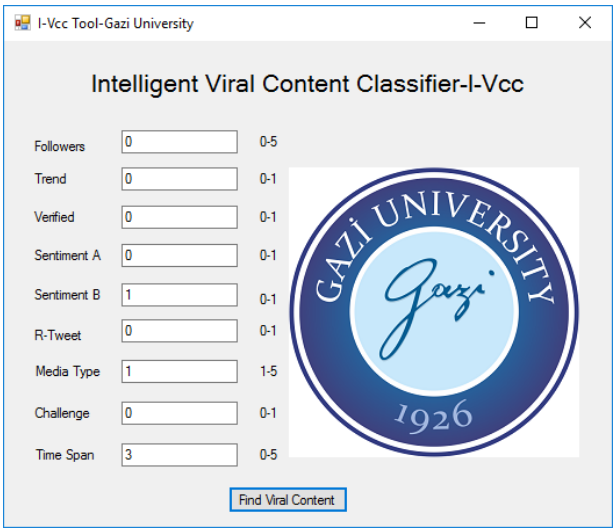

Fig. 5. I-VCC user interface

5. Conclusion

In this work, we proposed a methodology to classify the viral contents. Six of the parameters that we used are similar parameters in previous works with some adjustments and we recommended two new parameters that does not well studied in literature. We recommended a brief model to detect if a content is spreading epidemically on the web or not. Overall success rate in our experiments is 90% using ten-fold cross validation technique.

This study opens up a lot of future work areas that can help with the provenance problem and the prediction systems for diffusion in both physical and social networks. Knowing the viral content beforehand can be used in advertisement, industry, politics or any end user that wants to reach a high amount of people.

Other than predicting the content virality, transforming a non-viral content to a viral one, or increase its virality rate, the amount of people it can reach is an achieavable after right parameter adjustments made and it is a future plan for us with this study as a base.

Another problem is virus spreading in physical networks [29]. Some of the parameters like timing, content and trust are platform-free and opens up new study areas. This study can help predicting virus diffusion on physical networks before it spreads to a large amount of nodes.

Dataset we used is retrieved from Twitter under supervision to determine the infromation diffusion is epidemic or not. We processed the raw data before labeling phase. Data is labeled before the ANN classifier phase and the best ANN classifier model is determined to work with our proposed I-VCC application.

This study proves that the viral content is not always a random outbreak which cannot be predicted before it happens because of the parameters we use proves that connections exist between timing – virality, sentiment – virality etc. Proving that if there is a tie exists, we can predict the viral content before the content sharing starts and this study is a milestone to a prediction study. It is important to know if a content is viral or not because of the time and resource it can save. Also this is related to the provenance problem in literature.

References

[1] Barbier, G., Feng, Z., Gundecha, P., & Liu, H. (2013). Provenance data in social media. Synthesis Lectures on Data Mining and Knowledge Discovery, 4(1), 1-84.

[2] Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). The role of social networks in information diffusion. In Proceedings of the 21st international conference on World Wide Web (pp. 519-528). ACM.

[3] Milgram, S. (1967). The small world problem. Psychology today, 2(1), 60-67.

[4] Wu, F., Huberman, B. A., Adamic, L. A., & Tyler, J. R. (2004). Information flow in social groups. Physica A: Statistical Mechanics and its Applications, 337(1), 327-335.

[5] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social media mining: an introduction. Cambridge University Press.

[6] Keeling, M. J., & Eames, K. T. (2005). Networks and epidemic models. Journal of the Royal Society Interface, 2(4), 295-307.

[7] Jin, Fang, et al. "Epidemiological modeling of news and rumors on Twitter. "Proceedings of the 7th Workshop on Social Network Mining and Analysis. ACM, 2013

[8] Yang, Jaewon, and Jure Leskovec. "Modeling information diffusion in implicit networks." Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010.

[9] Leskovec, Jure, Lada A. Adamic, and Bernardo A. Huberman. "The dynamics of viral marketing." ACM Transactions on the Web (TWEB)1.1 (2007): 5

[10] Bakshy, Eytan, et al. "The role of social networks in information diffusion. "Proceedings of the 21st international conference on World Wide Web. ACM, 2012

[11] Aral, Sinan, and Dylan Walker. "Identifying influential and susceptible members of social networks." Science 337.6092 (2012): 337-341.

[12] Guille, Adrien, and Hakim Hacid. "A predictive model for the temporal dynamics of information diffusion in online social networks." Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012.

[13] Md. S. Bayzid, A. Iqbal, C. S. Hyder, M. T. Irfan," Application of Artificial Neural Network in Social Computing in the Context of Third World Countries", 5th International Conference on Electrical and Computer Engineering, ICECE, 2008

[14] M. Thelwall, D. Wilkinso, S. Uppal, "Data Mining Emotion in Social Network Communication: Gender Differences in MySpace", Journal of the American Society for Information Science and Technology, 2010.

[15] Zhao, J., Lui, J. C., Towsley, D., Guan, X., & Zhou, Y. (2011, April). Empirical analysis of the evolution of follower network: A case study on Douban. In*Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on* (pp. 924-929). IEEE.

[16] Ghenname, M., Abik, M., Ajhoun, R., Subercaze, J., Gravier, C., & Laforest, F. (2013, November). Personalized recommendation based hashtags on e-learning systems. In ISKO-Maghreb, 2013 3rd International Symposium (pp. 1-8). IEEE.

[17] Li, L. H., Lee, F. M., & Liu, W. J. (2006). The timely product recommendation based on RFM method.

[18] Shani, G., Meisles, A., Gleyzer, Y., Rokach, L., & Ben-Shimon, D. (2007). A stereotypes-based hybrid recommender system for media items. In Workshop on Intelligent Techniques for Web Personalization, Vancouver, Canada (pp. 76-83).

[19] Diao, Q., Qiu, M., Wu, C. Y., Smola, A. J., Jiang, J., & Wang, C. (2014, August). Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 193-202). ACM.

[20] Kong, D., & Zhai, Y. (2012, November). Trust based recommendation system in service-oriented cloud computing. In Cloud and Service Computing (CSC), 2012 International Conference on (pp. 176-179). IEEE.

[21] E. Tyugu, "Artificial Intelligence in Cyber Defense", 3rd International Conference on Cyber Conflict, pp. 1-11, Tallinn, 2011

[22] W. Britt, S. Gopalaswamy, J. A. Hamilton, G. V. Dozier, K. H. Chang, "Computer Defense Using Artificial Intelligence", Spring simulation multiconference, Vol. 3, pp. 378-386, 2007

[23] L.A. M. Pereira, L. C. S. Afonso, J. P. Papa, Z. A. Vale, C. C. O. Ramos, D.S. Gastaldello, A.N. Souza , "Multilayer Perceptron Neural Networks Training Through Charged System Search and its Application for Non-Technical Losses Detection", Innovative Smart Grid Technologies Latin America (ISGT LA), pp. 1-6, Sao Paulo, 2013

[24] Y. Lin, "Application of Extracted Rules from a Multilayer Perceptron Network to Moulding Machine Cycle Time Improvement", IEEE Transactions On Components, Packaging And Manufacturing Technology, Vol. 1(3), pp. 436 – 445, 2011

[25] S. Sagiroglu, U. Yavanoglu, E. N. Guven, "Web based Machine Learning for Language Identification and Translation", International Conference on Machine Learning and Applications, pp. 280-285, Cincinnati, OH, 2007

[26] U. Yavanoglu, O. Kaplan, H. Atli, G. Tanis, O. Milletsever, S. Sagiroglu, "Intelligent Decision Support System For Energy Investments", 12th International Conference on Machine Learning and Applications (ICMLA), vol. 2, pp. 224-231, Miami, FL, 2013

[27] Van Mieghem, P., Omic, J., & Kooij, R. (2009). Virus spread in networks.IEEE/ACM Transactions on Networking, 17(1), 1-14.