TELKOMNIKA, Vol.17, No.5, October 2019, pp.2667~2674
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i5.12646

The classification of the modern arabic poetry using machine learning

Munef Abdullah Ahmed*[1], Raed Abdulkareem Hasan[2], Ahmed Hussein Ali[3], Mostafa Abdulghafoor Mohammed[4]
[1]Faculty of Al Hawija Technical institute, Northern Technical University, Mosel, 41002, Iraq
[2]Faculty of Al-dour Technical institute, Northern Technical University, Mosel, 41002, Iraq
[3]AL Salam University College Computer Science Department Baghdad, Iraq
[4]Great Imam University College, Baghdad, 10053, Iraq
[4]Faculty of Automatic Control and Computers, University Polytechnic of Bucharest, 060042, Romania
*Corresponding author, e-mail: drmunef69@gmail.com[1], raed.isc.sa@gmail.com[2]
msc.ahmed.h.ali@gmail.com[3], alqaisy86@gmail.com[4]

Abstract

In recent years, working on text classification and analysis of Arabic texts using machine learning has seen some progress, but most of this research has not focused on Arabic poetry. Because of some difficulties in the analysis of Arabic poetry, it was required the use of standard Arabic language on which "Al Arud", the science of studying poetry is based. This paper presents an approach that uses machine learning for the classification of modern Arabic poetry into four types: love poems, Islamic poems, social poems, and political poems. Each of these species usually has features that indicate the class of the poem. Despite the challenges generated by the difficulty of the rules of the Arabic language on which this classification depends, we proposed a new automatic way of modern Arabic poems classification to solve these issues. The recommended method is suitable for the above-mentioned classes of poems. This study used Naïve Bayes, Support Vector Machines, and Linear Support Vector for the classification processes. Data preprocessing was an important step of the approach in this paper, as it increased the accuracy of the classification.

Keywords*: classification of arabic poems, machine learning algorithms, modern arabic poems*

1. Introduction

Despite the number of approaches on the automatic classification of the English language and other languages, the Arabic language still needs a lot of research, especially related to Arabic poetry. This is due to the number of determinants in the language, including its difficulty and the need to master the rules of the language when studying poetry. There is also a need for a full understanding of the theory of "Al Arud", which specializes in the study of Arabic poetry [1] whether as a regular text or poem, focused on the topic or on the effects [2]. Few studies have used sentiment analysis to classify Arabic texts [3]. In this study, we used Naïve Bayes (NB), Support Vector Machines (SVM), and Linear Support Vector classification (SVC) for the classification task.

The next section of this paper covers a review of the related work, followed by the introduction of the four categories of modern Arabic poetry. After that, the dataset of the work is presented, followed by the data preprocessing step which has a direct effect on the accuracy of the classification process. The sixth and seventh sections focus on feature selection and the machine learning algorithms used. These sections are followed by those that discuss the methodology, results, and conclusions from the study.

2. State of the ART

Several methods have been used in the English language for the classification of emotions. Some of these studies depended on keywords spotting or unambiguous words like "happy" and "sad" [4]. The lexical affinity from the effective research in this field depended on the emotion of the arbitrary term or words. In general, this method is better than the keyword

Received January 5, 2019; Revised March 17, 2019; Accepted April 22, 2019

spotting method as it cannot be used as an independent model [5]. There are other methods which rely on a deep understanding of the language and semantics [5]. Reliance on psychological theory in determining desires, goals, and needs was one of the models used in the classification [6]. The machine learning techniques used in the classification of classical Arabic poetry depended on the emotion [7]. This work classified the Arabic poetry into Fakhr, Retha, Ghazal, and Heija. The polynomial networks were used in the Arabic text classification [8]. Several classification algorithms have been used in the classification of Arabic text, such as SVM [8, 9], the NB [10], K-Nearest Neighbor (KNN) [11], Artificial Neural Network (ANN) [12], and the Rocchio feedback algorithm [13].

3. Categories of Modern Arabic Poetry

The modern Arabic poetry in general consists of the following types [14]:

- Love poems: It is a poetic art used to express the feelings between lovers. The poet derives the meanings of his relationship with the subject, his outlook, the influence of the environment, and the reality of those feelings.
- Islamic (religious) poems: The poets benefited from the stories contained in the Holy Quran; so, they took the precepts, rulings, and semantics and employed them in their poetry, treating community issues and problems that spread in their country at the time.
- Social poems: Social poems aim to repair bad social conditions by diagnosing the problem, identifying its cause, and describing its resolution. The poets resort to the method of encouragement and motivation when they want their people to contribute to the promotion and progress and avoid the pests and conditions that undermine the foundations of its renaissance.
- Political poems: This type of poetry expresses certain political orientations and the personal views of poets while preserving the way poetry is written, the values of literary and artistic poetry.

4. The Dataset

The Arabic language research using Natural Language Processing (NLP) is different from the English language in terms of the number and size of the datasets used. Due to the limited number of free available datasets in the Arabic language (which is an obstacle in the way of researchers), most researchers rely on a collection of datasets taken from magazines, news stations, and websites. Some researchers depended on Saudi newspapers [11]. In the Arabic research, several schools of thought have classified the datasets into training and testing groups. In our work, the big problem is finding the datasets for tuning and testing because it is the first work on using machine learning for classifying the modern Arabic poetry. We depended on the website for datasets to train and test the categories of modern Arabic poetry.

5. Data Pre-Processing

The Arabic language is difficult both in speaking and writing. It consists of 29 letters (أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي) and the ”Hamza” (ء) which are divided into two types. The first type is called long vowels, which includes three letters (ا, و, ي); the other is called constant letters. In this language, there are several kinds of diacritics used, such as “sukoon”, “dammah”, “Kasra”, “Fatha”, “tanween fatha”, “tanween kasra”, “tanween dammah”, “shadde”, and “mad”. These short vowels give correct pronunciation and meaning. Table 1 illustrates the short vowels and pronunciations to the words that have the same letters but different pronunciation and meaning as shown in Table 2.

Arabic writings are different from those using the Latin alphabet, due to the direction of writing from right to left. Some letters in Arabic also take several forms depending on the location of the character on the word. These features must be considered in this work as shown in Table 3.

Table 1. The Diacritics in Modern Arabic Poem

The short vowel	The Sign	Applied to the letter	Pronunciation
"sukoon"	ْ	لْ- سْ	S- L
"dammah"	ُ	لُ- سُ	Su - Lu
"Kasra"	ِ	لِ - سِ	Si - Li
"Fatha"	َ	لَ - سَ	Sa - La
"tanween fatha"	ً	لً - سً	San - Lan
"tanween kasra"	ٍ	لٍ - سٍ	Sin - Lin
"tanween dammah"	ٌ	لٌ - سٌ	Son - Lon
"shadde"	ّ	لّ — سّ	Ss – Ll
Mad	~	آ	Aa

Table 2. Example for the Effect the Diacritics on the Arabic Word

The word	The meaning
سَلَمَ	Hello
سُلَّمٌ	Ladder
سُلِم	Was delivered
سِلْم	Safety
سَلِمَ	Saved

Table 3. The Effect of a Positioning on the form of a Letter

The letter	The Arabic word	The meaning
هـ	هديه	Gift
ـهـ	الهام	Important
ـه	له	For him
ه	كره	A ball

The Arabic language has two types of genres, masculine and feminine. Each type in the Arabic language has different qualities and features in Arabic grammar. There are three classes in the Arabic language, the first is singular, the second is dual, and plural which also has two types (regular and broken). The Arabic language contains many ramifications in grammar. It is a very rich language, and this makes it difficult and a challenge to reach the required accuracy in the classification of modern Arabic poetry.

Pre-processing of data is an important thing to do when building classification systems using machine language for the following reasons:

- It removes noise from the text used in the classification.
- It reduces the terms or characteristics on which we base our classification.
- It helps reducing the amount of memory required for the classification.
- It helps increasing the accuracy of the classification.

We applied the following pre-processing on the data used in our work:

- Tokenization: We divided the data into parts and based on characteristics and recognition of delimiters like the punctuation of special characters and white space.
- We removed non-Arabic terms, words, numbers, punctuations, and any other singe.
- The stop words like pronouns, prepositions, and conjunctions were also removed; we deepened the list adopted by Khoja and Garside [15, 16].
- Stemming: The major aim of stemming is to decrease an inflated dataset. In Arabic, many words can be composed from the same stem. Thus, we can reduce the number of terms used in the dataset and the complexity of text classification. This is also a storage requirement for classification systems [17, 18].

6. Features Selection

In machine learning, constructing or representing vectors of features is a very important and critical point and has a significant impact on the results of the machine learning algorithm. Each object should be represented with its own features.

$$D = d1d2 \dots dn. \quad (1)$$

$$di = W1W2 \dots Wn \quad (2)$$

$$\ddot{d} = g(d) \quad (3)$$

where *D* is a document, W is a word, and g is the function representing the relation between the domain of documents and features. g may be a linear or nonlinear equation. The number of

classes is represented by C and the number of features is represented by K. C*K is a feature vector length. We performed the mutually deducted occurrence as follows: $n_c = (f_i)$ represented the probability of occurrence of feature f_i in category or class c. Therefore, the mutually deducted count feature became as follows:

$$dn_c(f_i) = n_c(f_i) - n_d(f_i), \text{ where } d \neq c, \quad (4)$$

which refers to the number of appearances of any characteristic or feature in any category deducted from the number of appearances of the same characteristic in all other categories. The feature vector was used for building document D once. When found any feature, the Boolean flag was used. The Boolean vector model used in this type of classification is better than the count model [19, 20].

7. Machine Learning Algorithms

In our approach, three machine learning algorithms were selected for the classification of modern Arabic poetry. These algorithms have been proven successful in the classification of the English text. The first algorithm is Support Vector Machines, the second is Naïve Bayes, and the third is Linear Support Vector Classification. The datasets consist of four groups (folders): Islamic contains 23 files, Love contains 25 files, Politic contains 22 files, and Social contains 22 files, as illustrated in Table 4. Classifier performance is evaluated by computing its precision [21], recall [16], and f-measure [22].

Table 4. The Datasets for the Classification

The folder name	Number of files	Number of verses
Islamic	23	600
Love	25	600
Politic	22	500
social	22	550

7.1. Support Vector Machines

SVM is a computationally kernel-based algorithm for regression and binary data classification purposes [17, 18]. Based on the structural risk minimization theory, the SVM has been proven successful in solving both local minimum and high dimensionality problems. It has a better generalization performance compared to other ML methods such as ANNs [19, 20]. SVM has so far been excellent in solving several real-world data mining predictive problems like time series prediction, text categorization, image processing, and pattern recognition [21, 22]. Despite the remarkable achievements of the SVM, there are still certain drawbacks that need to be addressed, such as problems on the relationship of the statistical learning theory with other theoretical frameworks, big data processing, parameters selection, and the generalization ability of a given problem [23, 24]. With the rate of development of information systems, high-dimensional, dynamic and complex data are easily generated [25, 26].

7.2. Naïve Bayes

The NB method is a classification scheme which relies on the Bayes' theorem. This technique assumes the independence of its predictors. Simply, the NB classifier assumes that there is no relationship between the existence of certain features in a class and that of any other feature [27-30]. This theory was adopted in determining the class of the document on the following equation:

$$C^* = argmax_c P(c|d) \quad (5)$$

where c represents the class and d represent the document.

$$C^* = argmax_c P(d|c) * \frac{p(c)}{p(d)} \quad (6)$$

Because p(d) has no effect or role, the equations become:

$$C^* = argmax_c P(d|c) * p(c) \quad (7)$$

The important hypothesis in this algorithm is that each property or feature in the document does not depend on the other's features, and assumptions produce the following equation:

$$C^* = argmax_c P(d|c) \prod_i^n p(f_i/c) * p(c) \quad (8)$$

7.3. Linear Support Vector Classification

Linear SVC is a type of machine learning algorithms similar to the SVM. Some features of this algorithm are the flexibility in selection and loss of functions. It is suitable for a huge number of samples. From the testing of this model on data, researchers have found it using one-against-rest approach compared to SVM which uses one-against-one approach. This model is used in several applications like the classification of text documents using sparse features [22-24].

8. Methodology

Figure 1 presents the outline of our work. In the beginning, we choose the dataset used in our work; after that, we segmented it into words and all the steps of data preprocessing were applied, including features extraction. We used three machine learning algorithms (SVM, LSVC, and NB) in training and testing.

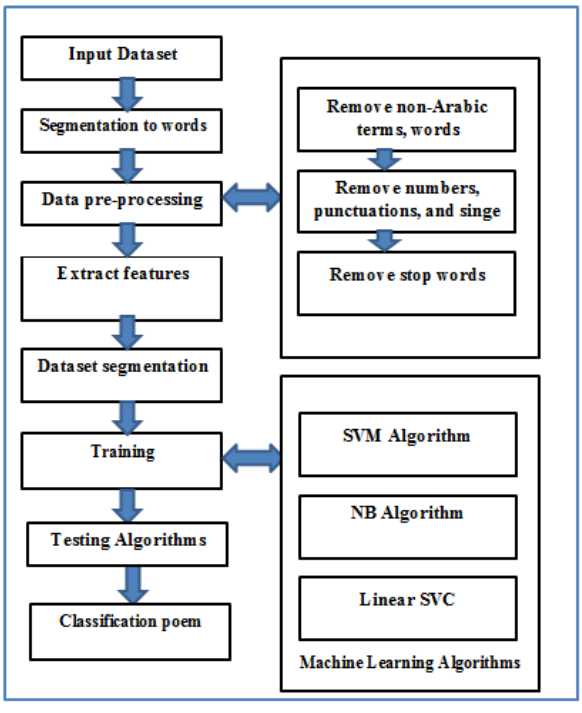

Figure 1. Block diagram of the proposed method

9. Results

The work was done with the Python language using the machine configuration as follows: OS: Windows 7, CPU Speed: 3.20 GHz, Processor: Intel Core i7, RAM: 4GB. With the intention of scrutinizing the suggested work's performance, different parameters such as precision, recall, and f-measure were measured for all types of modern Arabic poem.

The performance of the proposed method is presented in Tables 5 to 8 and Figures 2 to 5, as described below. The first type of machine learning algorithm used was Naïve Bayes. Table 5 illustrates the precision, recall, and f-measure for this algorithm. The maximum value for precision was for the politic class, while for the recall, the maximum value was for love class. F-measure was highest in the social and politic classes. The results for this algorithm were compared to the results of other machine learning algorithms.

Table 6 presents the results of the SVM algorithm. From the results, the maximum values of precision, recall, and f-measure were all for the Islamic class. This result was also compared to the results of the other machine learning frameworks. Table 7 illustrates the result of the classification process using linear SVC algorithm. From the results, the maximum value of precision was for the social class while the maximum values for recall and f-measure were for love class. Table 8 illustrates the average value for precision, recall, and f-measure for all the machine learning algorithms used in the classification of our dataset. From the table, linear SVC algorithm was found to have the maximum precision, recall, and f-measure values.

Figure 2 illustrates the precision for all types of modern Arabic poem using three machine learning algorithms. From the figure, the maximum value of precision for most types of the modern poem was presented by the linear SVC algorithm while the minimum value was presented by the SVM algorithm. When we compared the recall for our dataset as calculated using the tree machine learning algorithms, we found the maximum recall value in both NB and LSVC algorithms while the minimum recall value was found in SVM algorithm as shown in Figure 3. Figures 4 illustrates the f-measure for our dataset. The sequence of values from top to bottom in these algorithms was as follows: LSVC, NB, and SVM algorithm. Figure 5 illustrates the average value for our dataset. The best result was found in the L SVC algorithm, followed by the NB algorithm and SVM algorithm.

Table 5. Classification of our Dataset using Naïve Bayes

	precision	recall	F-measure
Islamic	0.14	0.5	0.22
Love	0.57	0.8	0.67
Politic	1	0.5	0.67
Social	0.5	0.17	0.25
Average	0.64	0.47	0.49

Table 6. Classification of our Dataset using Support Vector Machine

	precision	recall	F-measure
Islamic	0.5	0.25	0.33
Love	0.02	0.1	0.2
Politic	0.07	0.05	0.09
Social	0.12	0.16	0.1
Average	0.1775	0.14	0.18

Table 7. Classification of our Dataset using Linear Support Vector Classification

	precision	recall	F-measure
Islamic	0.17	0.5	0.25
Love	0.83	0.71	0.77
Politic	0.2	0.33	0.25
Social	1	0.29	0.44
Average	0.72	0.47	0.51

Table 8. Average results of our Dataset using Three Machine Learning Algorithms

	precision	recall	F-measure
Naïve Bayes	0.64	0.47	0.49
Support Vector Machine	0.1775	0.14	0.18
Linear Support Vector Classification	0.72	0.47	0.51

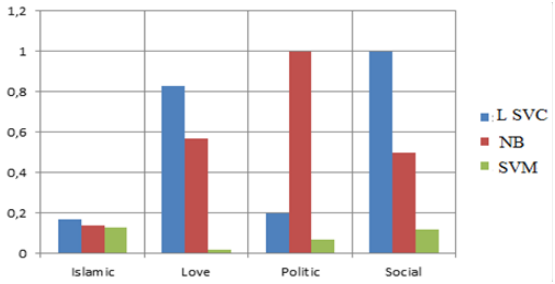

Figure 2. The precision for our dataset using three machine learning algorithms

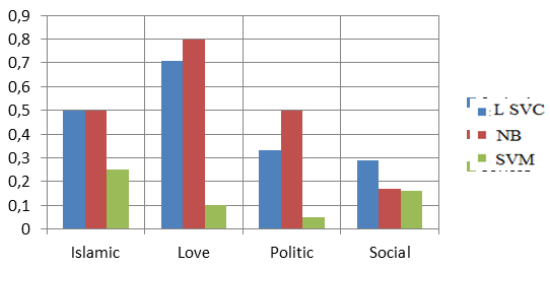

Figure 3. The recall for our dataset using three machine learning algorithms

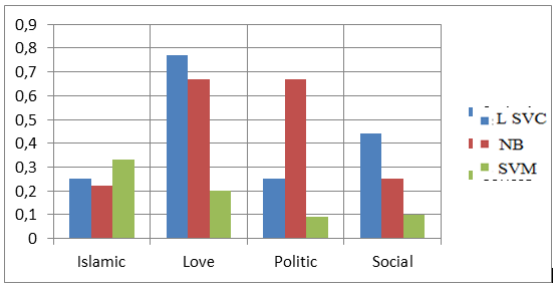

Figure 4. The F-measure for our dataset using three machine learning algorithms

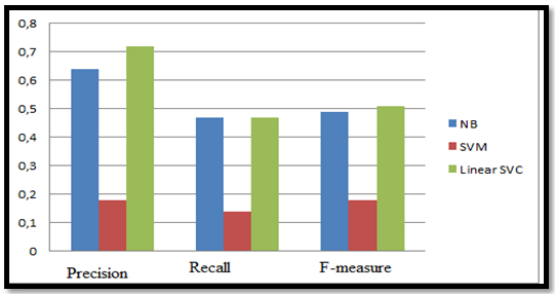

Figure 5. Average results of our dataset using three machine learning algorithms

10. Conclusion

In this paper, we used Support Vector Machine, linear Support Vector Classification, and Naïve Bayes for the classification of modern Arabic poems. The machine learning algorithms proved to be good tools for text classification. From the comparison of the result of the precision, recall, and f-measure for all types of the modern Arabic poem, the best result was found when using linear Support Vector Classification and Naïve Bayes. One of the main reasons for this disparity in performance could be the size of the dataset since some machine learning algorithms can work better with few datasets. Also, the preprocessing of our dataset was an important step as it increased the accuracy of the classification and reduced the required memory size for the classification process. This method of classification can be further improved for the other types of Arabic poetry.

References

[1] MA Ahmed, S Trausan-Matu. *Using natural language processing for analyzing Arabic poetry rhythm*. in Networking in Education and Research (RoEduNet), 2017 16th RoEduNet Conference. 2017: 1-5.
[2] S Al-Harbi, A Almuhareb, A Al-Thubaity, M Khorsheed, A Al-Rajeh. *Automatic Arabic text classification*. JADT 2008: 9es Journées internationales d'Analyse statistique des Données Textuelles. 2008: 77-83.
[3] M Abdul-Mageed, M T Diab, M Korayem. *Subjectivity and sentiment analysis of modern standard Arabic.* in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011; 2: 587-591.
[4] A Ortony, GL Clore, A Collins. The cognitive structure of emotions: Cambridge university press. 1990.
[5] H Liu, H Lieberman, T Selker. *A model of textual affect sensing using real-world knowledge*. in Proceedings of the 8th international conference on Intelligent user interfaces. 2003: 125-132.
[6] MG Dyer. Emotions and their computations: Three computer models. *Cognition and emotion*. 1987; 1(3): 323-347.
[7] O Alsharif, D Alshamaa, N Ghneim. Emotion classification in Arabic poetry using machine learning. *International Journal of Computer Applications*. 2013; 56(16):10-15.
[8] MM Al-Tahrawi, SN Al-Khatib. Arabic text classification using Polynomial Networks. *Journal of King Saud University-Computer and Information Sciences*. 2015; 27(4): 437-449.
[9] S Alsaleem. Automated Arabic Text Categorization using SVM and NB. in *Int. Arab J. e-Technol*. 2011; 2(2): 124-128.
[10] R Belkebir, A Guessoum. *A hybrid BSO-Chi2-SVM approach to Arabic text categorization*. in ACS International Conference on Computer Systems and Applications (AICCSA). 2013: 1-7.
[11] J Ababneh, O Almomani, W Hadi, NKT El-Omari, A Al-Ibrahim. Vector space models to classify Arabic text. *International Journal of Computer Trends and Technology (IJCTT)*. 2014; 7(4): 219-223.
[12] S Khorsheed, AOAl-Thubaity. Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language resources and evaluation*. 2013; 47(2): 513-538.
[13] L Fodil, H Sayoud, S Ouamour. Theme classification of Arabic text: A statistical approach. Terminology and Knowledge Engineering. 2014: 01005873.

[14] C Holes. Modern Arabic: Structures, functions, and varieties: Georgetown University Press. 2004.
[15] S Khoja, R Garside. Stemming arabic text. Lancaster, UK, Computing Department, Lancaster University. 1999.
[16] B Pang, L Lee, S Vaithyanathan. T*humbs up?: sentiment classification using machine learning techniques*. in Proceedings of the ACL-02 conference on Empirical methods in natural language processing. 2002; 10: 79-86.
[17] C Sudheer, R Maheswaran, B K Panigrahi, S Mathur. A hybrid SVM-PSO model for forecasting monthly streamflow. *Neural Computing and Applications*. 2014; 24(6): 1381-1389.
[18] X Zhang, S Ding, Y Xue. An improved multiple birth support vector machine for pattern classification. *Neurocomputing*. 2017; 225: 119-128.
[19] Z Chen, Z Qi, B Wang, L Cui, F Meng, Y Shi. Learning with label proportions based on nonparallel support vector machines. *Knowledge-Based Systems*. 2017; 119: 126-141.
[20] W Jiang, D-S Huang, S Li. Random walk-based solution to triple level stochastic point location problem. *IEEE transactions on cybernetic*s. 2016; 46(6): 1438-1451.
[21] T Joachims. *Text categorization with support vector machines: Learning with many relevant features*. in European conference on machine learning. 1998: 137-142.
[22] F Debole, F Sebastiani. An analysis of the relative hardness of Reuters-21578 subsets. *Journal of the Association for Information Science and Technology*. 2005; 56(6): 584-596.
[23] RA Hasan, MA Mohammed, ZH Salih, MAB Ameedeen, N Ţăpuş, MN Mohammed. HSO: A Hybrid Swarm Optimization Algorithm for Reducing Energy Consumption in the Cloudlets. *TELKOMNIKA Telecommunication, Computing, Electronics and Control*. 2018; 16(5): 2144-2154.
[24] RA Hasan, MA Mohammed, N Ţăpuş, OA Hammood. *A comprehensive study: Ant Colony Optimization (ACO) for facility layout problem*. in 2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet). 2017: 1-8.
[25] MA Mohammed, ZH Salih, N Ţăpuş, RAK Hasan. *Security and accountability for sharing the data stored in the cloud*. in 2016 15th RoEduNet Conference: Networking in Education and Research. 2016: 1-5.
[26] MA Mohammed, N ŢĂPUŞ. A Novel Approach of Reducing Energy Consumption by Utilizing Enthalpy in Mobile Cloud Computing. *Studies in Informatics and Control.* 2017; 26: 425-434.
[27] MA Mohammed, RA Hasan. *Particle swarm optimization for facility layout problems FLP—A comprehensive study*. in 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP). 2017: 93-99.
[28] ZH Salih, GT Hasan, MA Mohammed. *Investigate and analyze the levels of electromagnetic radiations emitted from underground power cables extended in modern cities*. in 2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2017.
[29] RA Hasan, MN Mohammed. A krill herd behaviour inspired load balancing of tasks in cloud computing. *Studies in Informatics and Control*. 2017; 26: 413-424.
[30] MA Mohammed, RA Hasan, MA Ahmed, N Tapus, MA Shanan, MK Khaleel, et al. *A Focal load balancer based algorithm for task assignment in cloud environment*. in 2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). 2018: 1-4.