Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM

ISSN: 2788-7421

A Classification of Al-hur Arabic Poetry and Classical Arabic Poetry by Using Support Vector Machine, Naïve Bayes, and Linear Support Vector Classification

Muneef Abdullah Ahmed[1,2,*]

[1]Electronics Department - Alhawija Technical Institute, Northern Technical University, IRAQ,
[2]Faculty of Automatic Control and Computers, University Polytechnic of Bucharest 313 Splaiul Independentei, 060042, Romania

*Corresponding Author: Muneef Abdullah Ahmed

DOI: https://doi.org/10.52866/ijcsm.2022.02.01.014
Received March2022; Accepted May 2022; Available online July 2022

ABSTRACT: Most of the world languages have made strides in analyzing and classifying texts electronically; hence, the use of electronic text has become a great alternative to manual classification as it reduces time, cost, and difficulty. However, in the Arabic language, electronic analysis has not progressed due to several limitations faced by researchers in this field, such as the complexity of the Arabic language, the lack of related research, as well as the use of the classical Arabic language. In addition, Arabic poetry has other limitations, such as the use of a system that uses a single activation function. In this research, a new method was developed for the classification of the classical Arabic poetry and Al-hur poetry. This new approach is based on features that indicate the type of poetry. Pre-processing of some data is important in this new approach as it helps increase the accuracy of classification.

Keywords: Automl, classification, arabic language classification.

1. INTRODUCTION

Despite the numerous languages that use Latin letters and their presence in all areas of language, whether poetic or literary, the Arabic language is lagging in this field. Hence, more research is needed, especially with regard to Arabic poetry, whether it is Al-hur Arabic poetry or classical Arabic poetry [1]. Research in this field requires extensive knowledge of all the rules of the Arabic language, its ramifications [2], and a full understanding of the theory of propositions for the Arab world. Furthermore [3], these studies rely on verbal content and not on emotions or feelings even though some studies relied on feelings in classifying the Arabic text by using machine learning (ML) [4]. The key steps in data pre-processing are:

- Tokenization: In this segment, the data are partitioned into parts on the basis of qualities and perceiving delimiters like punctuation and blank area.
- Elimination of non-Arabic numbers, words, terms, signs, and punctuations
- Removal of stop words (prepositions, pronouns, conjunctions), Khoja, and Garside [5].

ML impacts our lives and has a key role in all fields, whether it is in natural language processing, medicine, industries, or agriculture. Classification is one of the most important uses of ML. Support vector machine (SVM) is mostly used for classification and regression tasks due to its effectiveness and accuracy when using classified data [6].

This paper makes three key contributions:

- ML algorithms were used to classify classical Arabic poetry and Al-hur Arabic poetry.
- The two kinds of Arabic poetry were best compared by using Linear Support Vector Classification.

*Corresponding author: muneef_hwj@ntu.edu.iq

- The performance of Naïve Bayes and Support Vector Machine in classifying classical Arabic poetry and Al-hur Arabic poetry were determined.

This paper is organized in sections, beginning with the related work, followed by the types of Arabic poetry (classical Arabic Poetry and Al-hur Arabic poetry). The next section presents the datasets used, followed by the explanation of the ML algorithms used in this work. The methodology section was followed by the results and discussion section.

2. STATE OF THE ART

Many strategies have been utilized in the English language to determine the impact of classification [7]. A portion of these examinations relied upon watchword spotting or word without vagueness, like "good" and "bad" [8]. The lexical fondness from the viable examination in this field relied upon the feeling of the erratic term or words. This technique is superior to the catchphrase spotting strategy as it cannot be utilized as an autonomous model [9] Different techniques depend on a profound comprehension of the language and semantics. Dependence on mental hypothesis in deciding longings, objectives, and necessities was one of the models utilized in the grouping [10]. The AI utilized in the arrangement of classical Arabic poetry verse relied upon the inclination, and this work divided classical Arabic poetry into four types [11]. Polynomial organizations were utilized in Arabic text characterization [12]. Some algorithms have been utilized in Arabic text classification, like SVM [13], artificial neural network [14], Rocchio algorithm [15], k-nearest neighbors [16], and Naive Bayes [17].

3. CATEGORIES OF ARABIC POETRY

Arabic poetry can be classified into two types:

3.1 CLASSICAL ARABIC POETRY:

The verse or "byte" consists of two parts: the first part is called the "sadr" and the other part is called "ajuz." The last word in the "sadr" is called "arud" and the last word in the "ajuz" is called "hashu sadr." Moreover, the last word in "ajuz" is called "darb" and the other is called "hashu ajuz." Classical Arabic poetry has three styles [18]:

First style: In this style, the verse has one row and two columns; the first column is for "sadr" and the other is for "ajuz." See Figure 1 for an example.

Verse	H2 or "ajuz"	H1 or "sadr"
	وطبْ نفساً إذا حَكَمَ ألقضَاءُ	دَع الأيَّامَ تَفعلُ مَا تَشَاءُ

FIGURE 1. **First style of classical Arabic poem with one verse**

Second style: This style has two rows and one column; the first row is for "sadr" and the second row is for the "ajuz." See Figure 2 for an example.

Verse	H1 or "sader"	دَع الأيَّامَ تَـفـعَـلُ مَـا تَـشَـاءُ
	H2 or "ajuz"	و طبْ نَـفْـسـاً إذا حَـكَـمَ الـقَـضَـاءُ

FIGURE 2. **Second style of classical Arabic poem with one verse**

Third style: This style also has one column and two rows, but the first row is aligned to the right to represent the "sadr," whereas the second row, which represents the "ajuz," is aligned to the left. See Figure 3 for an example.

Al-Khalil bin Ahmed Al Farahidi is among the famous Arab scholars who focused on Arabic poetry and reached important milestones in this field. He laid the foundations of classical Arabic poetry as he is considered one of those who invented "buhur" of the classical Arabic poetry. He set the unit of measurement for these "buhur," which is called "tafilah." He concluded that classical Arabic poetry does not exceed 16 "buhur," namely, Muqtadib, Madid, Mutadariq,

Verse	H1 or “sadr”	دَع الأيَامَ تَفعلُ مَا تَشَاءْ
	H2 or “ajuz”	وطبْ نفسْاً إذا حَكَمَ ألقضَاء

FIGURE 3. **Third style of classical Arabic poem with one verse**

Tawil, Mutaqarib, Kamil, Mujtath, Basit, Munsarih, Wafir, Mudari, Hazaj, Sari, Hamal, Khafif, and Rajaz. These depend on the mobile letter or stationary letter; any changes in the cases of the letters can change the “buhur” for this verse into another “buhur.” Thus, all verses in a classical Arabic poem have the same “buhur.”

3.2 AL-HUR ARABIC POETRY:

The topics of Al-hur poetry focus on public life, such as social problems due to wars, illness, or distress. Al-hur poetry can also feature the poet’s expression of his feelings, the revolution against unjust rulers, the poet’s failure to protect his people, and the falsehood or injustice prevailing in a place. It consists of one line, which represents the “sadr” without “ajuz” (see Figure 4). It also has one “tafileh,” which is called “free,” because it is free of rhyme and form. The poet is free to vary the “tafileh” and length, but if the poem is organized on a specific “buhur,” then all the verses are on the same “buhur” [19].

Verses	Verse one “sader”	عيناكِ غابتا نخيل ساعةَ السحَرْ
	Verse two “sader”	أو شُرفتانِ راح ينأى عنهما القمر

FIGURE 4. **Style of Al-hur Arabic poems with two verses**

4. FEATURES OF AL-HUR POETRY

According to the rules of Al-Khalil bin Ahmed Al Farahidi, the same “tafileh” is used in Al-hur poetry, but the numbers of “tafileh” in all verses are not the same.

Recycling: An imperfect “tafileh” can be used at the end of the first verse and completed at the beginning of the second verse.

Organic unity: The organic consistency between the poem’s words, events, emotions, images, as well as the music is consistent with the emotions.

Ease of **language**: The emergence of the vernacular language makes the poem easy to understand due to the use of simple synonyms.

Realistic: As in ancient poetry, the truth and social problems are discussed.

5. DATASET

The use of a natural language processor in the Arabic language is considered difficult compared to the English language due to the lack of free data that researchers can use. Therefore, researchers depended on data taken from sources like magazines, websites, and news stations. Hence, the researchers differed in their views about the selection of the data set for training and testing, which is the first step on which ML depends. This is due to the lack of research in this field, especially in the field of Arabic poetry, whether it is classical Arabic poetry or free poetry.

6. DATA PRE-PROCESSING

The Arabic language is one of the most important global languages, and it consists of 29 letters (أ ب ج د هـ و ز ح ط ي ك ل م ن س ع ف ص ق ر ش ت ث خ ذ ض ظ غ) and Hamza (ء). Among the Arabic letters are three vowels letters (ا, و, ي), and the rest are constant letters. Table 1 shows the equivalent of Arabic vowels and letters in the English alphabet. The Arabic language is unlike other languages as it has several types of diacritics that may change the meaning and pronunciation. Table 2 specifies these diacritics in the Arabic language.

The Arabic language is written from right to left in contrast to the Latin language, Arabic contains letters that take several forms depending on their position in the word. This feature adds another difficulty to this language, which must be considered in this research. Table 3 shows how different positions of a letter can result in different words.

Table 1. E1: Equivalent of Arabic letters and vowels in the English alphabet

Arabic Letters		Equivalent Vowels	
ta, ta marbuta	ت, ة	/a/, /a:/	ا, ى, َ, ً
ha, ta marbuta	ه, ة	/u/, /u:/	و, وا, ُ, ٌ
		/i/, /i:/	ي, ِ, ٍ

Table 2. Types of diacritics in classical and Al-hur Arabic poems

Short vowel	Sign	Applied to the letter	Pronunciation
"sukoon"	◌ْ	رْ - زْ	R- Z
"dammah"	◌ُ	رُ – زُ	Ru - Zu
"Kasra"	◌ِ	رِ - زِ	Ri - Zi
"Fatha"	◌َ	رَ - زَ	Ra - Za
"tanween fatha"	◌ً	رً - زً	Ran - Zan
"tanween kasra"	◌ٍ	رٍ - زٍ	Rin - Zin
"tanween dammah"	◌ٌ	رٌ - زٌ	Ron - Zon
"shadde"	◌ّ	رّ – زّ	Rr – Zz
Mad	~	آ	Aa

Table 3. Different words formed by changing the position of a letter

Letter	Arabic word	Meaning	
"sukoon"	◌ْ	رْ - زْ	R- Z
"dammah"	◌ُ	رُ – زُ	Ru - Zu
"Kasra"	◌ِ	رِ - زِ	Ri - Zi
"Fatha"	◌َ	رَ - زَ	Ra - Za
"tanween fatha"	◌ً	رً - زً	Ran - Zan
"tanween kasra"	◌ٍ	رٍ - زٍ	Rin - Zin
"tanween dammah"	◌ٌ	رٌ - زٌ	Ron - Zon
"shadde"	◌ّ	رّ – زّ	Rr – Zz
Mad	~	آ	Aa

The Arabic language takes into account gender, that is, masculine or feminine, and each has its own rules. This language consists of three categories. In addition to the singular and plural forms, it contains the dual form. Thus, each gender has different rules in the singular, dual, and plural forms. To create a classification system by using ML, several steps must be done before processing to reduce noise and to determine or reduce the characteristics on which the classification depends. This helps reduce the memory used for classification in the event of large data, as well as increase the accuracy and speed of classification. The important steps that must be done during data pre-processing are:

Tokenization: In this segment, the data are partitioned into parts on the basis of the qualities and perceiving delimiters like punctuation and blank areas.

Elimination of non-Arabic numbers, words, terms, signs, and punctuations

Removal of stop words (prepositions, pronouns, conjunctions), Khoja, and Garside [5].

Stemming: The primary course for stemming is to reduce an expanded dataset. Many Arabic words can be formed from a similar stem; subsequently, the quantity of terms can be reduced by using the quality in the dataset and the intricacy of text arrangement. This step is additionally a capacity prerequisite for the arrangement of frameworks [20].

In ML, developing or addressing the vector of elements is a vital and fundamental step that affects the results of the AI calculation. Each item must be addressed with its own elements.

$$D = d1d2d3.........dn,$$

where the document is represented by D.

$$d = w1w2w3.........wn,$$

where the document is represented by w.

$$d_{=}g(d),$$

where g is the relation between the features and domain of D which may be linear or not linear; C and K represent the number of classes and the features for the vector, respectively; whereas the length of features vector is represented by C*K; the probability for each feature in the types of classes (c) can be written as follows: nc=(fi). Consequently, the commonly considered highlight becomes:

$$dn_c(Fi) = n_e(Fi) - n_d(Fi), whered \neq c,$$

which refers to the quantity of appearance of any features or element in any classification deducted from the quantity of appearances of similar features in any remaining classification [claas]. There are two types of vector models: Boolean and Count. This work used Boolean because it is more suitable than Count [21].

6.1 ALGORITHMS

Three types of ML were used in this work: SVM, Linear Support Vector Classification, and Naïve Bayes. These algorithms are mostly used for English text classification and have proven to be effective in classifying the English language. Our work consists of two folders; the first folder is the classical Arabic poetry which contains 30 files, and the other folder contains 33 files. See Table 4 for the details of the dataset.

Table 4. Datasets for the classification

Name of the folder	Number of files	Number of verses
Classical Arabic poetry	30	800
Al-hur Arabic poetry	33	780

6.2 SUPPORT VECTOR MACHINE

The world of ML is considered one of the important worlds that has an impact on our real life and in all fields, whether it is in natural language processing, medical, industrial, or agricultural. Classification is one of the most important uses of ML. SVM is mostly used for classification and regression tasks due to its effectiveness and accuracy when using classified data [22]. Classification has two types: binary classification classifies two groups, whereas multiple classification classifies several groups. This algorithm is popular mainly in the field of diagnosing patterns, images processing, and distinguishing handwritten numbers [23]. If the data are simple, then linear classification can be used. However, if the data are complex and contain several dimensions, then the kernel technique of the algorithm is used.

Classification accuracy is estimated by calculating tree features, namely, precision, recall, and F1-measure.
The precision for class ci can be obtained as follows [21]:

$$Pi = \frac{TPi}{(TPi + FPi)} \tag{1}$$

Recall for class ci is calculated as:

$$Ri = \frac{TPi}{(TPi + FNi)}, \tag{2}$$

where:
Tpi refers to true positive;
FNi refers to false negative;
FPi refers to false positive.
Van Rijsbergen calculated the F-measure by using the following function [22]:

$$F1 = (2 * Recall * Precision/(Recall + Precision)); \tag{3}$$

$$= (2TP/(2TP + Fp + FN)). \tag{4}$$

6.3 LINEAR SUPPORT VECTOR CLASSIFICATION

Linear Support Vector Classification is considered one of the important ML algorithms similar to SVM. However, it differs in the form of its execution as it is executed in liblinear and not libsvm. One of the key characteristics of this algorithm is the ability to choose and lose functions; it is also used in calculating the characteristics of large numbers of samples. Through several testing, studies have found that SVM is based on a one-against-one approach, whereas the LSVC is based on the one-against-rest approach. It is used in several applications, but it is distinguished in the field of classification of natural languages [24].

6.4 NAÏVE BAYES

This ML algorithm depends on Bayes' theorem that assumes no relationship between specific traits in a class and any other traits. In this theory, the class of documents is given as:

$$C* = argmax_c P(c/d),$$

where c and d represent class and document, respectively.

$$C* = argmax_c P(d/c) * P(c)/P(d) \tag{5}$$

According to this theory, p (d) does not have an effect, and the equations will be as follows:

$$C* = argmax_c P(d/c) * p(c). \tag{6}$$

In this theory, the feature does not depend on the other feature; therefore:

$$C* = argmax_c P(d/c)\pi_i^n p(f_i/c) * p(c) \tag{7}$$

7. METHODOLOGY

Figure 1 illustrates the block diagram for our method of classification, which includes several steps.

8. RESULTS

For the SVM algorithms, the maximum precision and recall for the classical Arabic poetry and maximum F-measure for Al-hur Arabic Poetry are presented in Table 5.

For the Naïve Bayes algorithm, the classical Arabic poetry has maximum precision and recall, whereas Al-hur Arabic poetry has a maximum F-measure, as shown in Table 6.

The maximum precision and F-measure for the SVC algorithm can be seen in the classical Arabic poetry, whereas the Al-hur Arabic poetry has the maximun recall as presented in Table 7.

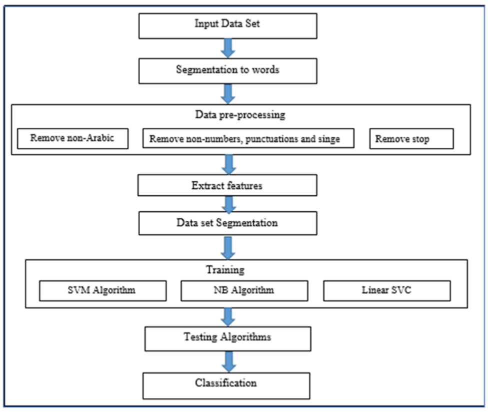

FIGURE 5. Block diagram of the proposed method of classification

Table 5. Support Vector Machine Classification

	Precision	Recall	F-measure
Classical Arabic poetry	0.6	0.32	0.4
Al-hur Arabic poetry	0.03	0.1	0.35
Average	0.315	0.21	0.375

Table 6. Naïve Bayes Classification

	Precision	Recall	F-measure
Classical Arabic poetry	0.5	0.9	0.58
Al-hur Arabic poetry	0.9	0.6	0.67
Average	0.7	0.75	0.625

Table 7. Classification of Linear Support Vector

	Precision	Recall	F-measure
Classical Arabic poetry	0.83	0.53	0.77
Al-hur Arabic poetry	0.67	0.7	0.56
Average	0.75	0.615	0.665

Table 8. Average Classification for Algorithms

	Precision	Recall	F-measure
Support Vector Machine	0.315	0.21	0.375
Naïve Bayes	0.7	0.75	0.625
Linear Support Vector Classification	0.75	0.615	0.665

The maximum average precision and F-measure were obtained by the LSVC algorithm, whereas the NB achieved the maximum average recall, as shown in Table 8.

Figure 2 illustrates the precision for all the ML algorithms used. The maximum precision was achieved by the NB algorithm for Al-hur Arabic poetry, whereas the minimum precision for Al-hur Arabic poetry was achieved by SVM algorithm.

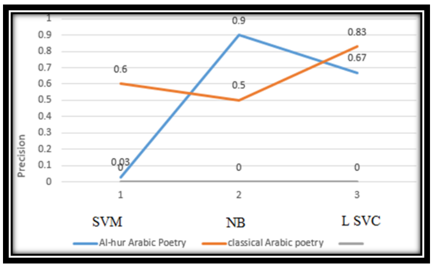

FIGURE 6. Precision for classical and Al-hur Arabic poetry in the utilized ML algorithms

Figure 3 presents the recall for the utilized ML algorithms. The comparison shows that the maximum recall value for classical Arabic poetry was achieved by NB algorithm, whereas the minimum recall value for Al-hur Arabic poetry was achieved by SVM algorithm.

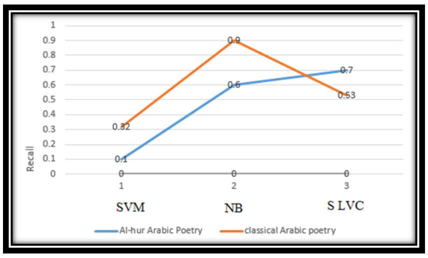

FIGURE 7. Recall for classical and Al-hur Arabic poetry for the utilized ML algorithms

Figure 4 presents the F-measure for all the used ML algorithms. The comparison shows that the maximum F-measure value for the classical Arabic poetry was achieved by LSVC, whereas the minimum F-measure value for the Al-hur Arabic poetry was achieved by SVM algorithms.

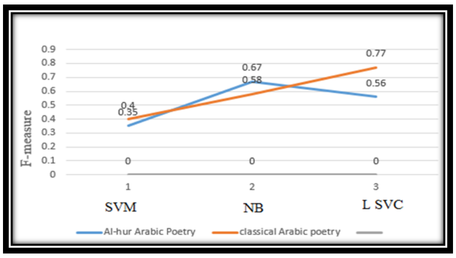

FIGURE 8. F-measure for classical and Al-hur Arabic poetry for the utilized ML algorithms

Figure 5 illustrates the average of precision, F-measure, and recall values for classical and Al-hur Arabic poetry for all the utilized ML algorithms. The precision and F-measure values were maximum when the LSVC algorithms were used, whereas the maximum recall value was obtained with the NB algorithm.

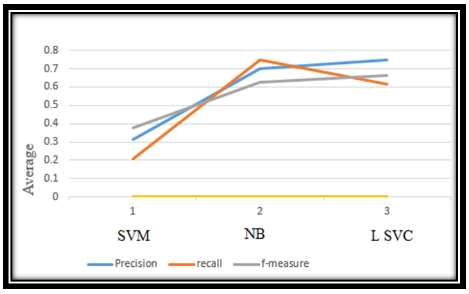

FIGURE 9. Average precision, recall, and F-measure values for classical and Al-hur Arabic poetry for the utilized ML algorithms

9. CONCLUSION

ML algorithms are mostly used for classification tasks, and they are excellent for classifying languages with Latin letters. In this paper, ML algorithms were used to classify classical Arabic poetry and Al-hur Arabic poetry. The results showed that the best precision, F-measure, and recall for the utilized ML algorithms on the two kinds of Arabic poetry were achieved by using Linear Support Vector Classification, followed by Naïve Bayes, and lastly by SVM. The reason for this difference in performance is the size of the data used as some ML algorithms perform better on large datasets. Furthermore, one of the factors that affect the accuracy of the results is the preprocessing of the data used in the classification task.

ACKNOWLEDGEMENT

The first author would like to thank the reviewers for providing useful suggestions, allowing for the improved presentation of this paper.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] M. Ahmed, A. H. Hasan, and Ali *MA Mohammed The classification of the modern Arabic poetry using machine learning Telkomnika*, vol. 17, pp. 2667–2674.
[2] M. A. Ahmed and S. Trausan-Matu *Using natural language processing for analyzing Arabic poetry rhythm*, pp. 1–5.
[3] M. Abdul-Mageed, M. T. Diab, and M. Korayem *Subjectivity and sentiment analysis of modern standard Arabic*, pp. 587–591.
[4] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. Khorsheed, and A. Al-Rajeh *Automatic Arabic text classification*, 2008.
[5] G. L. Ortony, A. Clore, and Collins *The cognitive structure of emotions*, 1990.
[6] H. Liu, H. Lieberman, and T. Selker *A model of textual affect sensing using real-world knowledge*, pp. 125–132.
[7] O. Alsharif, D. Alshamaa, and N. Ghneim, "Emotion classification in Arabic poetry using machine learning," *International Journal of Computer Applications*, vol. 65, no. 16, 2013.
[8] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB," *Int. Arab J. e-Technol*, vol. 2, no. 2, pp. 124–128, 2011.
[9] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB," *Int. Arab J. e-Technol*, vol. 2, no. 2, pp. 124–128, 2011.
[10] R. Belkebir and A. Guessoum *A hybrid BSO-Chi2-SVM approach to Arabic text categorization*, pp. 1–7.
[11] M. S. Khorsheed and A. O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," *Language resources and evaluation*, vol. 47, pp. 513–538, 2013.
[12] M. M. Al-Tahrawi and S. N. Al-Khatib, "Arabic text classification using Polynomial Networks," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 4, pp. 437–449, 2015.
[13] J. Ababneh, O. Almomani, W. Hadi, N. K. T. El-Omari, and A. Al-Ibrahim, "Vector space models to classify Arabic text," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 7, no. 4, pp. 219–223, 2014.
[14] M. S. Khorsheed and A. O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic dataset," *Language resources and evaluation*, vol. 47, pp. 513–538, 2013.
[15] L. Fodil, H. Sayoud, and S. Ouamour *Theme classification of Arabic text: A statistical approach*, vol. 10.
[16] Holes *Modern Arabic: Structures, functions, and varieties*, 2004.
[17] S. Khoja and R. Garside, *Stemming arabic text*. 1999.
[18] L. Pang, S. Lee, and Vaithyanathan *Thumbs up?: sentiment classification using machine learning techniques*, pp. 79–86.
[19] X. Zhang, S. Ding, and Y. Xue, "An improved multiple birth support vector machine for pattern classification," *Neurocomputing*, vol. 225, pp. 119–128, 2017.
[20] T. Joachims *Text categorization with support vector machines: Learning with many relevant features*, pp. 137–142.
[21] W. Jiang, D. S. Huang, and S. Li, "Random walk-based solution to triple level stochastic point location problem," *IEEE transactions on cybernetics*, vol. 46, no. 6, pp. 1438–1451, 2016.

[22] F. Debole and F. Sebastiani, “An analysis of the relative hardness of Reuters-21578 subsets,” *Journal of the Association for Information Science and Technology*, vol. 56, no. 6, pp. 584–596, 2005.

[23] M. A. Mohammed, A. A. Kamil, R. A. Hasan, and N. Tapus, “An Effective Context Sensitive Offloading System for Mobile Cloud Environments using Support Value-based Classification,” *Scalable Computing: Practice and Experience*, vol. 20, no. 4, pp. 687–698, 2019.

[24] R. A. Hasan, S. N. Shahab, and M. A. Ahmed, “Correlation with the fundamental PSO and PSO modifications to be hybrid swarm optimization,” *Iraqi Journal For Computer Science and Mathematics*, vol. 2, no. 2, pp. 25–32, 2021.