Indonesian Journal of Electrical Engineering and Computer Science
Vol. 16, No. 3, December 2019, pp. 1482~1489
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v16.i3.pp1482-1489

A framework for sentiment analysis in Arabic text

Alaa Abdalqahar Jihad [1], Ahmed Subhi Abdalkafor [2]
[1]Computer Center, University of Anbar, Iraq
[2]Career Development Center, University of Anbar, Iraq

Article Info

Article history:

Received Feb 26, 2019
Revised Jun 1, 2019
Accepted Jul 22, 2019

Keywords:

Arabic text classification
Bag of words
Natural language processing
Sentiment analysis
Text mining

ABSTRACT

Over the last decade there has been an increase in number of E-mails or comments to a company via social media sites, to satisfy their customers, the company must take in to consideration these messages and comments and know whether the customers are satisfied with what the company offers or not. Several techniques have been proposed to analyze the sentiment of the comment writer. Dealing with the Arabic language is faced with many challenges, such as it is a morphologically rich language and how to return the word to its original root. In this paper the challenges of dealing with the Arabic language were reviewed and a framework was also established to analyze the comments in Arabic and classify it into positive, negative or neutral sentiment. The framework was trained and tested and then the conclusions were drawn based on its work.

Corresponding Author:

Alaa Abdalqahar Jihad,
Computer Center,
University of Anbar, Anbar, Iraq.
Email: it.alaa.heety@uoanbar.edu.iq

1. INTRODUCTION

The flow of information over the Internet is too large to search for an automatic analysis of very important documents and texts [1]. An analysis of feelings or opinions is the use of natural language processing, textual analysis and computer linguistics for the purpose of detecting feelings that are useful or neutral towards the subject of the text. Another concept is to identify the sensory tone of a series of words for the purpose of understanding opinions and emotions, whether they are sad or happy. Emotional analysis generally aims to identify the feelings of a speaker or writer about a subject or to identify the predominant feelings of a document writer. These feelings can express the author's opinion or his emotional state [2], the data is classified into positive, negative or mixed As well as the tendencies of the person in terms of psychological of depression and anxiety [3-4]. Emotion analysis is used in areas of marketing, customer service, and other areas.

The organizational processes of knowledge management systems such as academic libraries through the classification of scientific publications, which help researchers to find useful information and quick access to useful articles among millions of articles. There are many studies and practical application to the analysis of the text, or classification on English language and other languages [5-7], several studies have been suggested in Sentiment Analysis of Arabic Text, the contributions to the existing studies are the following:

Mohammad et al [8]. The study included a discussion of the problem of classifying the Arabic text using three algorithms: Naïve Bayes (NB), Support vector machine (SVM) and Neural Network (NN) and then applied a comparative study on a large Arabic database. In this study, a steady number of Arabic documents were used in the training phase. The study included several stages. First, Preprocessing the purpose of this stage is, to easily handle docu-ments and reduce complications such as deleting unnecessary words such as stop words and tags. This process is followed by the representation of the document where it is converted to the vector area and then the reduction of dimensions is chosen so that the text becomes free from the complexities and ready to extract properties, this process includes cleaning the text and stemming to be presented in clear format and then identify the most important features appropriate to the original text. After conducting the training and testing of three algorithms and comparing them. The results showed that the SVM algorithm supports the best results.

Al-Anzi and AbuZeina [9]. In this study, the technique of cosine similarity was used for the purpose of verifying the performance of the classification of the Arabic text. The single value analysis method (SVD) was used to index the underlying significance (LSI) to extract the textual properties. The SLI technique is to represent the text in a better way because it maintains the semantic information between the words. A number of classification methods have been applied (Classification Tree (CT), Neural Network (NN), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN)) over 4,000 documents for 10 topics (400 documents per topic). In this study, the method (TF.IDF) was used to extract features. After analysis and testing, the results showed that the use of (LSI) is superior to (TF.IDF) and the K-Nearest Neighbors (KNN) method of classification is the best among the methods of classification used in this study.

Bahassine et al [1]. This proposed study included two main aspects; the first aspect included the development of new algorithms to represent each term of a particular document at its root. The other side of this study was the study of the comparison of the two algorithms (new stemmer and Khoja’s stemmer) on 5070 documents classified independently into six categories (entertainment, Middle East, sport, switch and world on WEKA toolkit a business). Precision measures and f-measure were used to compare the performance of models. The experimental results of this study showed the superiority of the proposed algorithm to classify the text; it reached 89.1% in the business category and 92.9% in the sports category.

Al-Sabahi et al [10]. In this study, a method of semantic analysis of Arabic documents was improved, where statistical and linear methods were used to overcome restrictions that reduce performance. A part of the speech clarification tool was used to minimize dimensions (LSA). For the purpose of considering the order of words and grammatical relationships during the calculation of the matrix, the weight of the term was added in four sentences to weighting schemes. Also, to make the summary that was created more useful, the description of the term and the description of the sentence were combined for each subject. An extensive trial of four sets of data has been applied to both English and Arabic to ensure the effectiveness of the proposed algorithm. The results showed the best results compared to the latest technology.

Alowaidi et al [11]. In this study, a model has developed a model for a semantic Arabic text based on the application of Twitter using semantic analysis and machine learning. The synonyms that appear in the Twitter application are represented as different independent features. To overcome these limitations, the tweets were represented by an external knowledge base (Arabic WordNet (AWN)). In this study, methods of representing and evaluating different concepts were developed using algorithms SVM and NB classifiers. The experimental results showed an improvement in the performance of the proposed model compared to the basic model where the ratios were reached 5.78% in NB classifier while the ratio reached 4.48% in SVM classifier.

Froud and Ouatik [12]. This paper aims to automatically compile similar documents into one cluster. This paper aims to automatically collect similar documents in one document cluster and reduce the noise in the document information to enhance the performance of documents clustering. In this study, the effect of summarizing the text was evaluated using the latent analysis model for the underlying significance of the compilation Arabic documents using the following methods: Jaccard Coefficient (JC), Euclidean Distance (ED), Pearson Correlation Coefficient (PCC), Cosine Similarity (CS) and Averaged Kullback-Leibler Divergence (AKD). The experimental results showed that the proposed solutions solve the problems of the length of documents as well as the noise information thus a high improvement in the performance of documents clustering.

Bilal and Rasha [13]. In this paper, a group of opinions on restaurants and the conclusion of linguistic characteristics was analyzed to be used in the similarity measures. The most important features were extracted based on the knowledge base and the distributional similarity between the aspects of revisions and specifications. The results showed high performance of the application after applying this proposal to help academic challenges dataset. There is very little of this research headed towards the Arabic language and therefore there are several reasons for this, including the rules of Arabic language and the formation of letters where they are connected and not separate, one punch, each letter of this language enters into several primitive, intermediate, finite and separated groups as well there are different forms of words.

1.1. Applications of Text Analysis

Massive amount of data is increasing day by day so we need to maintain and analyze data for effective process. There is great importance in the processing of text mining and higher commercial potential of data extraction, text analysis can be used in many applications including [14-19]:

1) Classification of texts to specific ranges.
2) Sentiment analysis.
3) Summarize the document to provide the most important points in the original document.
4) Learn about relationships between specific entities.
5) Search for information in the text.
6) Categories to make them easier for potential users to identify.
7) Organize texts so that the user can easily access them.
8) The opinions can be analyzed for obtaining consideration of the decision-making in any organization.

1.2. Arabic Challenges

Dealing with texts in Arabic includes many challenges [20-23] :

1) Arabic is morphologically rich language.
2) Return the word to its original root (Stemmer).
3) Contain words in colloquial terms.
4) Texts contain English words written in Arabic letters.
5) Writing characters in place of other characters.
6) Additions, prefixes, middle and endings.
7) The number of words is huge.
8) Spelling and grammatical errors in texts.
9) Highly use of Arabic pronouns.
10) Consonant doubled (ّ).

2. THE PROPOSED FRAMEWORK

The proposed framework consists of two parts; part for training and part for testing. The model used is bag of words and try to make them efficient method by adding weight for each word, this weight was added based on repeating the word in texts when training. What distinguishes the framework is that the database was built without stemmer, including a database of positive words and a database of negative words as in Table 1 and Table 2 show example on the stop and negation words.

Table 1. Example On The Positive and Negative Words

Positive		Negative	
Word	Weight	Word	Weight
زين	62	مش	57
الحمد	36	بس	39
خير	28	حرام	19
نعم	15	مو	16
جميل	13	وين	15
احلى	12	ليش	15
رائع	11	للأسف	11
صح	10	قليل	11
رائع	9	مشاكل	9
حلو	8	تكاليف	9

Table 2. Example on the stop and negation words

Stop words	Negation words
ان	**ليس**
بعد	**ليست**
الى	**غير**
في	**لا**
من	*(Colloquial term)* **مو**

Figure 1 illustrates the training steps in the framework, and Figure 2 shows the steps of testing and analyzing the comments.

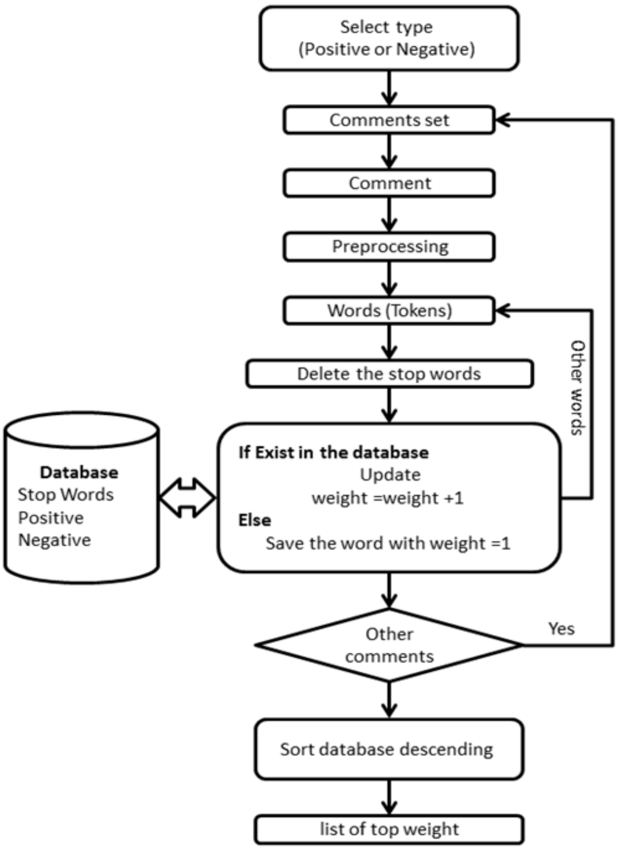


Figure 1. Training steps

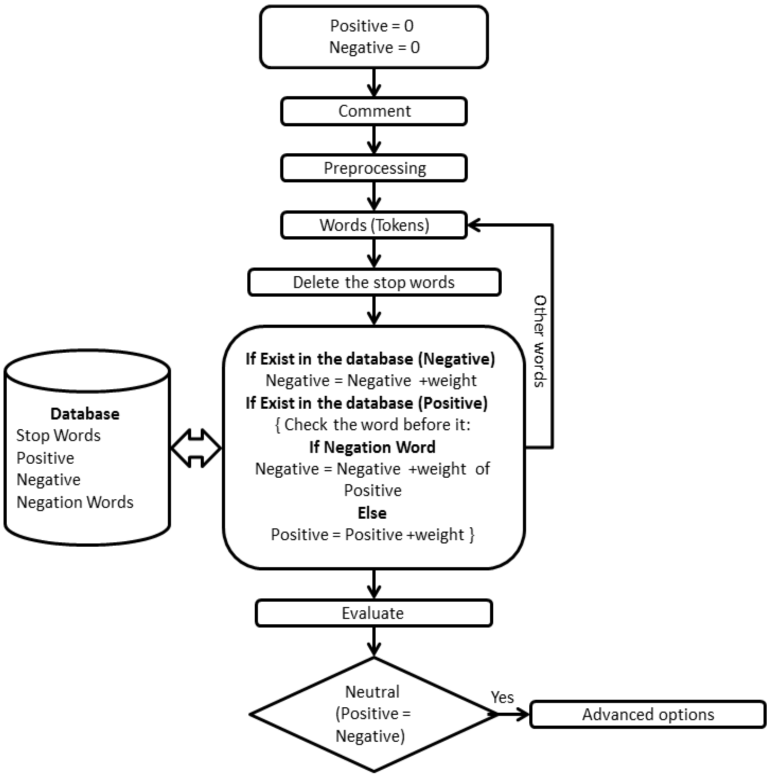

Positive = 0
Negative = 0
Comment
Preprocessing
Words (Tokens)
Delete the stop words
Database
Stop Words
Positive
Negative
Negation Words
If Exist in the database (Negative)
Negative = Negative +weight
If Exist in the database (Positive)
{ Check the word before it:
If Negation Word
Negative = Negative +weight of Positive
Else
Positive = Positive +weight }
Other words
Evaluate
Neutral (Positive = Negative)
Yes
Advanced options

Figure 2. Testing steps

2.1. Dataset Description

The dataset used from commercial site (for selling products) where the comment is written directly in the site or through social media sites or Email of the company.

2.2. Training

The following are the training steps, the input is a set of comments and output is a database containing negative and positive words with their weights:

1) Select type of sentiment (Positive or Negative) and its own Comments set.
2) Apply the following to each Comment in the Comments set:
   a. Bring the Comment.
   b. Preprocessing: deleting extra spaces, symbols and duplicate letters.
   c. Divide into words (Tokens) and store in an array.
   d. Delete the stop words.
   e. Save new words that are not in the database with a weight equal to 1.
   f. Or update the weight of the words in the database by one increment.
   g. Sort the database in a descending pattern based on the weight.
   h. Taking the beginning of the list that have top weight, and manually filtering it and save into a selected type of sentiment.

2.3. Testing

The following are the steps of testing any comment, the input is the comment and the output is evaluating the comment whether it is negative or positive or neutral:

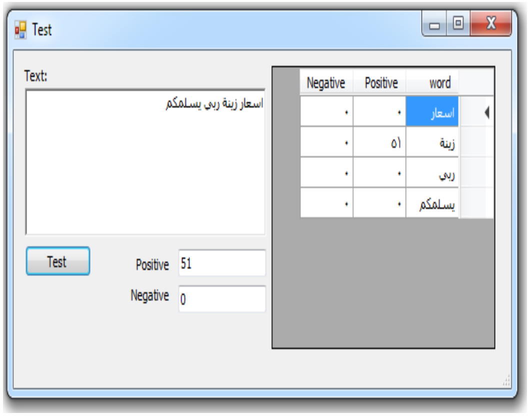
1) Let: Positive = 0 and Negative = 0.
2) Bring the Comment.
3) Preprocessing: deleting extra spaces, symbols and duplicate letters.
4) Divide into words (Tokens) and store in an array.
5) Delete the stop words.
6) For each word in the array:
   a. Search in the Negative words database, if the word exists, increase Negative by the weight of the word.
   b. Search in the Positive words database, if the word exists:
      - Check the word before it, if it is negation increase Negative by the weight of the Positive word.
      - Otherwise increase positive by the weight of the word.
7) Evaluating the result.

If the result is Neutral (Positive results= Negative results) show advanced options for manual classification if the user wants.

3. RESULTS AND DISCUSSION

The proposed framework was created using C# language. The test results of the framework were satisfactory, Figure 3 shows the comment test (اسعار زينة ربي يسلمكم) where the result (Positive = 51 and Negative = 0), the Figure 4 illustrates the example of the preprocessing and test step. When testing a comment (الجهاز غير جيد) in Figure 5 note that although there is a word (جيد) but the result was (Positive = 0 and Negative = 9) because there is a negative word (غير) before it. Figure 6 shows the comment test (أحتاج الى 1500 قطعة غيار للموديل) where the result (Positive = 0 and Negative = 0) is neutral so advanced user options have been shown.

The classification used in the analysis of feeling is divided into two types. The first one is the classification of data by subject to determine the accuracy of data both in training or in the testing phase, and the second type is the classification of data by the analysis of sentiment. The performance of the system has been achieved a high speed in implementation and high accuracy in results by using the first type of classification.



Test

Text:

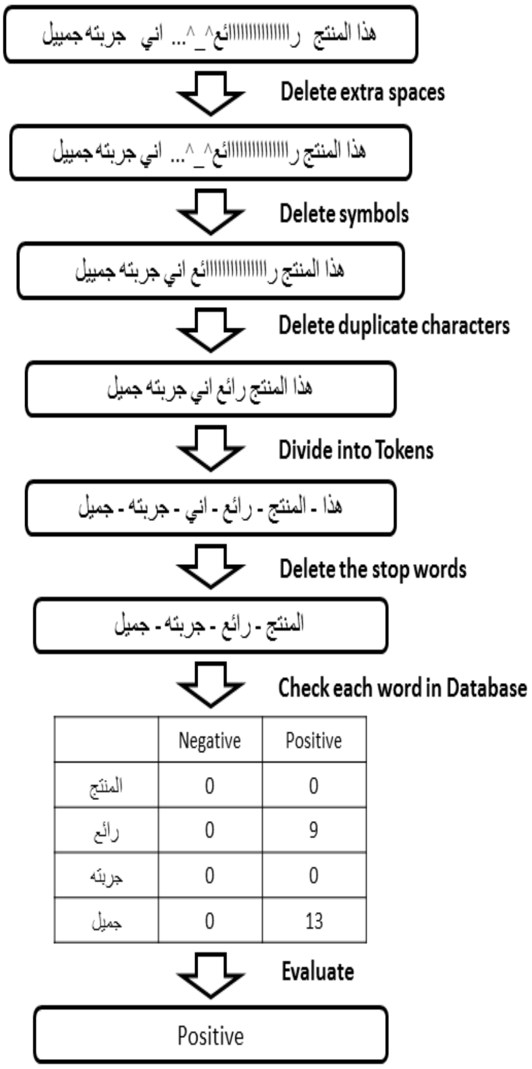
اسعار زينة ربي يسلمكم

word	Positive	Negative
اسعار	٠	٠
زينة	٥١	٠
ربي	٠	٠
يسلمكم	٠	٠

Positive: 51

Negative: 0

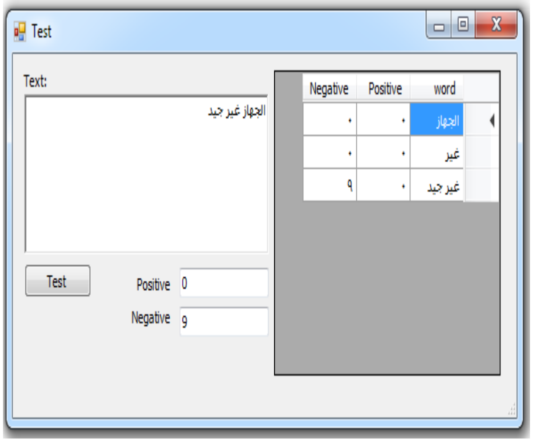
Figure 3. A comment Test



هذا المنتج   راااااااااااااائع^_^... اني   جربته جمييل
Delete extra spaces
هذا المنتج راااااااااااااائع^_^... اني جربته جمييل
Delete symbols
هذا المنتج راااااااااااااائع اني جربته جمييل
Delete duplicate characters
هذا المنتج رائع اني جربته جميل
Divide into Tokens
هذا - المنتج - رائع - اني - جربته - جميل
Delete the stop words
المنتج - رائع - جربته - جميل
Check each word in Database
Evaluate
Positive

	Negative	Positive
المنتج	0	0
رائع	0	9
جربته	0	0
جميل	0	13

Figure 4. Example of the preprocessing and test step



Test

Text:

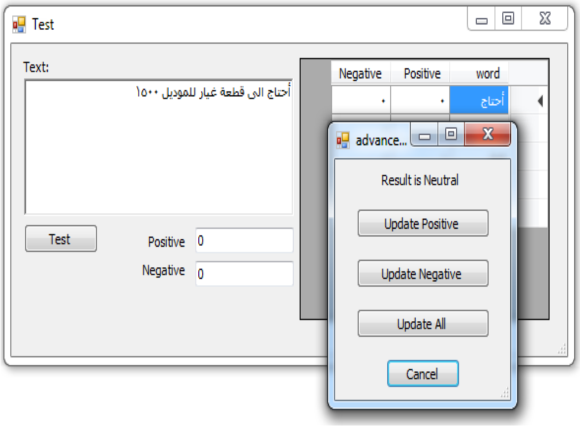
الجهاز غير جيد

word	Positive	Negative
الجهاز	٠	٠
غير	٠	٠
غير جيد	٠	٩

Positive: 0

Negative: 9

Figure 5. A test of a comment



Test

Text:

أحتاج الى قطعة غيار للموديل ١٥٠٠

word	Positive	Negative
أحتاج	٠	٠

Positive: 0

Negative: 0

advance...

Result is Neutral

Figure 6. A comment test and result is neutral

4. CONCLUSION

This paper deals with the comments in Arabic and the main difficulty in dealing with this language is that it is morphologically rich language; the company reaches (especially if it has a web site and social contact site) hundreds of comments and messages. The company can improve the performance through the development of systems to help the company to analyze customer requirements and analyzing the views of customers.

Using the bag word helped the accuracy of the classification. Enable the user using advanced options when the test result appears 'Neutral' helped to strengthen the database of words and appropriate weights, at the same time this feature is optional. Entering the largest amount of data in the training stage can improve the performance and achieve high accuracy results but lower in the execution time.

REFERENCES

[1] Bahassine S., et al., "Arabic text classification using new stemmer for feature selection and decision trees," *Journal of Engineering Science and Technology*, vol. 12, pp. 1475-1487, 2017.

[2] Froud H., et al., "Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering," *arXiv preprint arXiv:1302.1612*, 2013.

[3] H. G. Hassan, et al., "A Framework for Arabic Concept-Level Sentiment Analysis using SenticNet," *International Journal of Electrical and Computer Engineering*, vol. 8, pp. 4015, 2018.

[4] S. Budiyanto, et al., "Depression and Anxiety Detection Through the Closed-loop Method using DASS-21," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 17, 2019.

[5] A. S. Abdalkafor, "Designing Offline Arabic Handwritten Isolated Character Recognition System using Artificial Neural Network Approach," *International Journal of Technology*, vol. 8, pp. 528-538, 2017.

[6] A. S. Abdalkafor and A. Sadeq, "Arabic Offline Handwritten Isolated Character Recognition System Using Neural Network," *International Journal of Business and ICT*, vol. 2, pp. 41-50, 2016.

[7] A. S. Abdalkafor, et al., "Predicting The Success Rates of Schools Using Artificial Neural Network," *Journal of Theoretical and Applied Information Technology*, vol. 96, pp. 6339-6348, 2018.

[8] Mohammad A. H., et al., "Arabic text categorization using support vector machine," *Naïve Bayes and neural network. GSTF Journal on Computing (JoC)*, vol. 5, 2018.

[9] Al-Anzi F. S. and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, pp. 189-195, 2017.

[10] K. Al-Sabahi, et al., "An Enhanced Latent Semantic Analysis Approach for Arabic Document Summarization," *Arabian Journal for Science and Engineering*, pp. 1-16, 2018.

[11] Alowaidi S., et al., "Semantic Sentiment Analysis of Arabic Texts," *International Journal of Advanced Computer Science and Applications*, vol. 8, pp. 256-262, 2017.

[12] Froud H., et al., "Arabic text summarization based on latent semantic analysis to enhance arabic documents clustering," *arXiv preprint arXiv:1302.1612*, 2013

[13] Bilal G. A. and N. Rasha, "Semantic Analysis based Customer Reviews Feature Extraction," *Journal of University of Babylon*, vol. 25, pp. 802-813, 2017.

[14] N. Ksh, et al., "Document representation techniques and their effect on the document Clustering and Classification: A Review," *International Journal of Advanced Research in Computer Science*, vol. 8, pp. 1780-1784, 2017.

[15] Ismail H., et al., "Automatic Arabic Text Categorisation: A Comprehensive Comparative Study," *Journal of Information Science*, vol. 41, pp. 114-11, 2015.

[16] F. Leila, et al., "Theme Classification of Arabic Text: A Statistical Approach," *Conference paper, Terminology and Knowledge Engineering, Berlin, Germany*, pp. 10, 2014.

[17] M. Fikri and R. A. Sarno, "Comparative study of sentiment analysis using SVM and SentiWordNet," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, pp. 902-909, 2019.

[18] A. S. Abdalkafor, et al., "A Novel Database for Arabic Handwritten Recognition (NDAHR) System," *2st International Conference on Computer Applications & Information Security (ICCAIS)*, 2019.

[19] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis–a hybrid scheme," *Journal of Information Science*, vol. 42, pp. 782-797, 2016.

[20] Milos R. and Mirjana I., "Text Mining: Approaches And Applications," *Novi Sad J. Math*, vol. 38, pp. 227-234, 2008.

[21] A. S. Abdalkafor, "Survey for Databases on Arabic Off-line Handwritten Characters Recognition System," *1st International Conference on Computer Applications & Information Security (ICCAIS), IEEE*, pp. 1-6, 2018.

[22] M. A. Ahmed, et al., "The classification of the modern arabic poetry using machine learning," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 17, pp. 2667-2674, 2019.

[23] A. M. F. Al Sbou, "A Survey of Arabic Text Classification Models," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, pp. 4352-4355, 2018.

BIOGRAPHIES OF AUTHORS

Alaa Abdalqahar Jihad was born in Anbar-Iraq in 1985. He received his B.Sc. from Faculty of Computer Science at Anbar University, Iraq in 2009. The MSc. degree Faculty of Computer Science at Anbar University, Iraq 2012. His research interests are, Data Warehouse, Data Mining, Artificial Intelligent, Machine Learning and Natural Language Processing.

Ahmed Subhi Abdalkafor was born in Anbar-Iraq in 1988. He received his B.Sc. from Faculty of Computer Science at Anbar University, Iraq in 2010. The MSc. degree from Computer Science Department in Middle East University, Jordan in 2016. His research interests are, Image Processing, Pattern Classification, Artificial Intelligent, Neural Network, Cloud Computing, Machine Learning and Data Mining.