JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-15, No.-10, October (2020) pp 20-34 ISSN (Print) 0973-8975

AUTOMATIC ARABIC KEYWORD EXTRACTION USING LOGISTIC REGRESSION

Noor T. Jabury[1], Nada A.Z. Abdullah[2]

[1,2]Department of Computer Science, University of Baghdad, Baghdad, Iraq

Corresponding Author: **Noor Tahseen Jabury**

Email: nadaazah@scbaghdad.edu.iq

https://doi.org/10.26782/jmcms.2020.10.00002

(Received: August 17, 2020; Accepted: October 1, 2020)

Abstract

Keywords Express the main content of the document or article, they are an important component since they provide a summary of the article's content. Keywords also play an important role in information retrieval systems, bibliographic databases, and search engine optimization. The manual assignment of high-quality keywords is expensive, time-consuming, and error-prone. In this paper, an automatic keyword extraction model, based on the Logistic Regression algorithm is proposed and implemented. The model consists of three main stages: preprocessing, feature extraction, and classification stage to select the keywords. In experimental results 40 Arabic documents are used from two Arabic journals (AJSP and JJSS), the results are promising; the average accuracy is 0.91 with average precision 0.86 for the AJSP dataset, the average accuracy is 0.90with average precision 0.83 for the JJSS dataset.

Keywords : Arabic keywords, keywords extraction, logistic regression

I. Introduction

One of the greatest significant tasks in text mining is Automatic Keyword Extraction (AKE) that changes the manual extraction of keywords measured as a time-intense procedure. English Keyword extraction is an old topic, which implemented using various methods. However, keyword extraction for Arabic documents is still a novel topic and has a small amount of research despite the significance of the Arabic language [X].

The Arabic language is one of the most important languages on the Internet, many Internet magazines and papers are in the Arabic language. Limited tries have been prepared to develop a keyword extraction from Arabic texts. [XXI]

The proposed is a model of Arabic keyword extraction from Arabic documents; this model contains three stages: the first stage is preprocessing to sift text from Punctuation marks, Symbols, and numbers. The second stage is feature extraction, in this work (eight) features are used; first is the Term Frequency (TF) procedure, which weights the words, others are related to the position of the word in the title, abstract,

and how the word is printed. The third stage is a classification of words using a logistics regression algorithm, which is one of the supervised learning algorithms.

The rest of this paper is as follows, the next section is a review of some researches related to automatic Arabic keyword extraction in previous years. In section 3, challenges in the Arabic language is discussed, section 4 presents the keyword extraction approaches, section 5 illustrates the proposed model stages. Afterward, section 6 presents an evaluation of the proposed approach and test results. Conclusion and future work, also presented at the end.

II. Related Work

The important information is indicated by the keyword extraction, that information is included in the document for enhancing the keyword execution. This section is including several recent studies that made for the keyword extraction techniques:

In 2013 Al-Kabi, M. et al. made a study about the system of keyword extraction that working on Arabic documents using the co-occurrence statistical information used in the English and Chinese language systems. The basic of the proposed method is the top frequent terms extraction and build the matrix of co-occurrence, it shows the frequent term occurrence. If the co-occurrence for a particular term is in the degree of business, therefore, that term is important and it mostly is a keyword. The term business degree and the frequent terms set are measured by using the $\chi 2$. As a result, the high $\chi 2$ values terms are mostly, be keywords. The method of $\chi 2$ that was adopted in this study compared with a different novel method using the frequency-inverted term and term frequency (TF-ITF) that first time tested. The number of datasets used was two for evaluating the performance of the system. The obtained results show that method of $\chi 2$ is better than the method of TF-ITF, the $\chi 2$ accuracy was 0.58 and the recall was 0.63, the second experiment of the $\chi 2$ obtained 64% of accuracy the two experiments results shows the ability of the $\chi 2$ method in applying it on the Arabic documents and its performance is acceptable compared to other techniques [XII].

In 2014 Awjan, Arafat Atwi et al. made a study that presents the unsupervised approach combined of two-phases for the keyword extraction from the document written in Arabic that combining the statistical analysis and linguistic information. The first phase is detecting all of the N-grams that could be taken as keywords. The second phase is analyses of the N-grams by the use of the morphological analyzer for replacing the N-grams words with their basic forms that consider the roots and body of the derived word of the nob-derivative words. The N-grams that contain a similar base form are regrouping with the accumulation of their counts. The proposed work results are achieved 0.51 of accuracy. The experiments of the proposed work are extracting the keywords from a single document in a way of domain-independent. The analysis text linguistic and the N-grams grouping according to their linguistic features is improving the extracted keywords quality [II].

In 2016 Omoush, Ebtehal H., et al. made a study using a self-organization map (SOM) neural network as a method of unsupervised learning. The proposed method performance is using the F-measure, recall, and precision for the evaluation. This technique is using two datasets the first is the JJSS dataset and the second is the

Wikipedia dataset. The obtained results show a precision of 42.84% by using the JJSS dataset and 46% by using the Wikipedia dataset [VIII].

In 2017 Suleiman, D. et al. made a study about a new method of keyword extraction by using the bag-of-concept for extracting the keywords from the Arabic text. The algorithm proposed is utilizing the model of semantic vector space instead of the traditional model of vector space to words group into classes. The word context matrix is built by the new method and the synonyms words are grouping in the same class. The new approach evaluations lead to the use of a dataset consisting of three documents and can be compared with the keywords extraction using the method of equivalence classes' term from the Arabic documents. The obtained results from the proposed method are 90% of precision in the second document and it is considered the best because the number of the keywords is small [VII].

In 2019 Armouty, B., & Tedmori, S et al. made a study about keyword extraction from an Arabic document. The proposed method used statistical features in a supervised learning technique and Support Vector Machine classifier. This method is applied to Arabic news documents. The result showed a precision of 0.77 and a recall of 0.58 [IV].

III. The Principle Challenges of Arabic Language

Many challenges in Arabic language related to the keyword extraction, the following are the main challenges: [XVI]

• Variations in Orthography:

There are specific Arabic characters combinations not unique in the rendition. For example, the symbols that combined the name HAMZA with the "أ" the name HAMZA dropping the "ا". This makes the symbol of whether HAMZA present.

• Complex Morphology:

The Arabic language has a high inflection degree. As an example gives the possessive, the word should have the letter "ي" called YA; it is connected to the end of the Arabic word. The disjoint does not exist in the Arabic language such as "MY".

• Broken Plurals:

The English language has broken plurals with a resemblance to a specific form that singular. The broken plurals in the Arabic language are not the same as the English language. The broken plurals in the Arabic language are not bounded to the morphological rules also is hard to be related to a singular form.

• Arabic Words are Derived:

The words in the Arabic language are often derived from a simple bare verb called root that containing three letters. One letter from the root could be dropped or sometimes more than one letter in some derivations. Sometimes the root tracing that is derived from a particular word can be a big problem.

• **Vowel-Diacritics**:

The vowels are often deleted from written Arabic. This deletion of that vowels leads to ambiguity through the interpretation of the words.

• **Synonyms**:

The synonyms are widely known and used in all kinds of Arabic literature. To make a better recall, the synonyms must be considered during the processing of the query.

IV. Keyword Extraction Approaches

The Pin-I and Shi-Jen [XV] is divided the approaches of automatic keyword extraction as Statistics Approaches and Machine Learning Approaches, Zhang et al. [X] explain it with more details, and they proposed four categories as the following:

- **Statistics Approaches**

It is including the simple methods that not requiring training data. The words statistics is using for identification of the keywords such as TFIDE, n-gram statistics, word co-occurrence, word frequency, PAT Tree (Patricia Tree, a suffix tree or position tree), etc. The disadvantage is that some professional texts like medical and health, the keyword can be shown in the article only one time. Using the statistically empowered models could filter the words by accident [VI].

- **Linguistics Approaches**

The linguistic feature is using the sentences and documents, words mainly. Semantic, lexical, discourse and syntactic analysis are famous and widely known analyses but it considers complex.

- **Machine Learning Approaches**

Considering the learning of supervised or unsupervised from examples, and the keyword extraction related work is preferring the approach of supervised. The approaches of supervised machine learning-induced a trained model upon a collection of keywords. Manual writing is required for learning datasets that are considered boring and inconsistent. The keywords unfortunately are assigned by the authors when they complete their documents. So, the generated model is applied for the keyword -extraction from a novel document. The approach is including an SVM, Bagging, Naïve Bayes, etc. the training data are required by a method and sometimes depends on the domain. The re-learning is needed from the system and after re-learning establishing the model whenever changing the domain the induced model is demanding very much and also consuming much time when the dataset is massive. [VIII]

- **Other Approaches**

The methods that the above mentioned are combined by the keyword extraction. Also, they are incorporating the heuristic knowledge for fusion sometimes, like the length, HTML, and similar tags, position, the text formatting, the layout of terms, etc. [XXV].

V. The Proposed AKE Model

The model of AKE is consisting of three phases: the first phase is the preprocessing phase; it is applying the tokenization, performing the normalization, and filter the tokens from stop word and stemming. The second phase is the feature extraction phase, the final phase is the decision making phase, every word in the word-list is applying the decision making phase on it for determining that word either a keyword or not a keyword. The proposed solution uses linear and logistic regression models learned from human-labeled data .the model of AKE consisting of major phases described in the following section. Figure 1 illustrates the proposed work main stages.

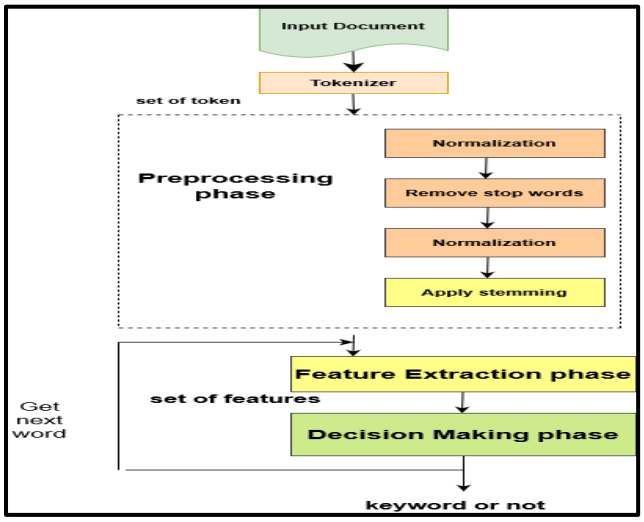

Fig. 1: General flowchart to describe the major stages of the proposed method

V.i. Preprocessing Phase

The processes of data preparing for the task of core text mining are called preprocessing. Those processes in converting the source of original data into a specific format to apply different methods of feature extraction on these documents for creating a new document collection that is represented by the concepts [IX]. The phase of preprocessing is including all processes, routines, and methods that are required for preparing the data used in the system of text mining that is considered the knowledge discovery operations core.

In the stage of preprocessing the text is split into tokens, this operation is called tokenized for creating a group of tokens and after that, the stop words of Arabic will be removed. The words like the preposition and pronouns are not useful in the categorization of text [XXV]. The preprocessing of data is using to remove the factors of language-dependent from the text and consists of the removal of stop words, tokenization, or stemming [XVIII]. The preprocessing of document or dimensionality reduction (DR) is allowing for skilled data manipulation from the text that is categorized. The dimensionality reduction is considering an essential step

during the process of classification. It is allowing the delete of the unimportant features of the document, hence, that leads sometimes to minimize the efficiency of the classification, also minimize their accuracy and speed [XXIV]. The preprocessing of the document is including 4 steps that are illustrated next subsection and in the following figure 2. Illustrates the preprocessing stage.

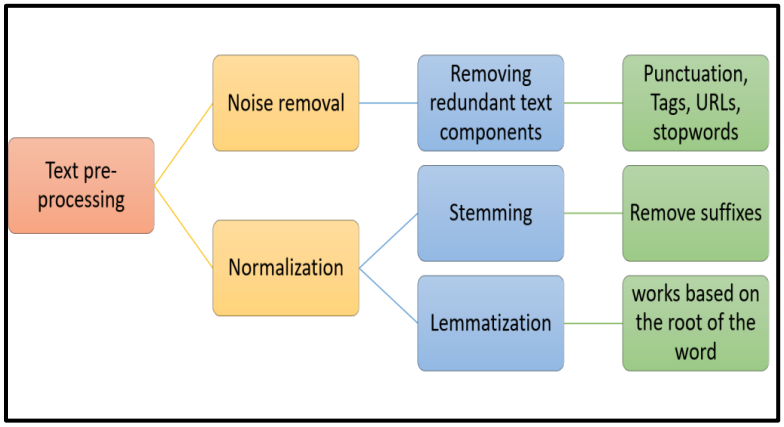

Fig. 2: Preprocessing Steps

Algorithm 1. Illustrates the preprocessing basic steps. The algorithm steps are used for converting the text from human language into a machine-readable format for more processing using the library of Natural Language Toolkit (NLTK).

Algorithm (1): The preprocessing Algorithm

Input document (T)

Stop word list (SWL)

Character & number list (CNL)

Output: a collection of tokens for the document(S)

Begin

Step 1: Read the document.

Step 2: Remove number from text

Step 3: Remove punctuation from text

Step 4: Break every word from others depending on space to acquire (Token).

Step 4: take out stopwords

Step 5: Stemming the residual token// extract the root of the tokens by using the ISRI stemming algorithm

Return (list of tokens)

End

V.ii. Features Extraction Phase

In the classification of text, the text feature extraction is playing an essential role; it is influenced in the text classification accuracy [XXII]. The process of feature extraction is reducing the dimensionality in a way of reduction of the initial set of raw data into manageable groups for processing. The characteristics of these sets are represented by a large number of variables that need many computing resources for processing. The name feature extraction is for the methods that selecting or combining the variables into features, it reduces effectively several data that needed to be processed, while it describes the original dataset completely and accurately. Eight features are used in this work which is summarized in table 1, the Features are:

- **Term Frequency**

The term frequency represents the measurement of the specific keyword appearing among a group of documents. It is called the keyword frequency also; the term frequency is the measurement of the key phrase or word repeating indicating how much that word is important to that document. The term frequency is rewarding the tokens that appear a lot of times in the same document. If any word occurring in a document often, then it

Considers an important to the document meaning, term frequency is the computation of the tokens in a document frequency. [V]

Eq. (1) is used to find TF.

$$\text{TF weight} = \frac{Mumber\ of\ times\ the\ token\ appears\ in\ the\ document}{Total\ number\ of\ token\ in\ the\ document} \quad (1)$$

- **Title Words**

The title considers an important part of the article, the researchers' investigation and selection of title have an important effect on the citation likelihood increasing of the article. So, when the main keywords in the article's titles are recognized, it represents the content indicative of that article and also helping to access to relevant articles easier by many people such as researchers. As a result, the article's title will be writing with more care from the authors.

This will not be only retrieving the articles of the authors' and researchers' faster, but also increasing these articles citations. Hence, the word inside the title will be taken into consideration. [XXII]

- **Abstract Words**

The Abstract represents the research article, review, or thesis summary, it is using for helping the readers to know the purpose of the article and a quick and fast way. The big amount of information available online is considered difficult to be handled. The abstract is helping in reducing this problem. The keywords may help more because it considers short summaries. The keywords used in document filtering while searching could save some time. The word appears in the abstract will be taken into consideration [XVII].

- **First Sentence (FS)**

Words that appear in the first sentence of the document can be important and are taken into consideration as a feature. This feature the word appeared in the first sentence will be set to 1 else will be set to 0 if it no appears in the first sentence.

- **Selection Based On Informative Features**

The words exist in the different writing forms in a particular document which will be providing additional information about the importance of the word. There are many different information features like the words emphasized by the application of underline fonts, bold font, or italic font [XIII].

- **Part of Speech**

The Part-of-Speech Tagger or it is known as POS Tagger represents a software piece that is used for reading the text in a particular language and assigns speech parts to every word and another token, as an example in the Arabic can be particle, verb or noun [XIV].

Table 1: Features used in the proposed model

No	Features	Descriptions	Regularization Method
1	TF	The part of the number of epochs tokens occurs in the text to a total of tokens in the documents.	$TF_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{i,j}}$
2	T	If the token has shown in the title	[0,1]
3	A	If the token has shown in the abstract	[0,1]
4	Fs	If the word has shown in the first sentence	[0,1]
5	B	If the token is bold	[0,1]
6	*I*	If the word is italic	[0,1]
7	U	If the word is underline	[0,1]
8	Pos tag	(Part-Of-Speech of token)if the token in an expression is N ,POS=1, else, POS=0	[0,1]

V.iii. Logistic Regression

It is a classification algorithm using for assigning observations to a discrete set of classes. The classification problem examples such as Email spam or not spam online transactions Fraud or not Fraud, Tumor Malignant, or Benign. The output of logistic regression is transformed by the use of the logistic sigmoid function for returning the value of probability [XX]. The logistic regression is providing the user with classification explicit probabilities away from the information of the class label, which is considered one of the advantages of logistic regression. Also, it is very easy to be extending to the classification problem of multi-category [XIX]. The data include two columns; (label and features) are predictable by the experiment as it is set up:

1-Label assigned to a row of training data. Here, labels are either "none keyword" or "keyword.

2-Features include the document to which the label applies. Algorithm 2 presents the steps of proposed AEK using the LR algorithm.

Algorithm 2 AKE using LR algorithm

Input: Features of the document tokens

Output: classification result keyword or not keyword

Begin

Step1: import preprocessing function

Step2: prepare feature extraction file

Step 4: splitting features in to training X and testing set Y

Train-test-split(x,y,size-test,random-state)

Step 5: Activation function used to map any real value between 0 and 1

Return 1 / (1 + np.exp(-x))

Step 6: Computes the cost function for all the training samples

m = x.shape[0]

*total_cost = -(1 / m) * np.sum(y * np.log(probability(theta, x)) + (1 - y) * np.log(*

1 - Probability (theta, x)))

Return total_cost

Step 7: Computes the gradient of the cost function at the point theta

m = x.shape[0]

*Return (1 / m) * np.dot(x.T, sigmoid(net_input(theta, x)) - y)*

Step8: train the model using the training set

Logistic_ regression. Fit(xtrain ,ytrain)

Step9: making prediction on the testing set

y-pred= Logistic_ regression.predict(x test)

Step10: comparing actual response value (y-test) with predicted value

End

VI. Performance Evaluation

To evaluate the proposed model two types of datasets are used the Arab Journal for Scientific Publishing (AJSP), the anther dataset from the Jordan Journal of Social Sciences (JJSS). Twenty academic papers are taken from each journal. We arbitrarily choose academic papers every text comprises the title, abstract, keywords full-document, borders information of sections, references. These texts have rich semantics topographies and are appropriate to achieve keywords classification in good form. Figure 3 shows one of the Arab Journal for Scientific Publishing paper.

Fig. 3: Arabic paper form Arab Journal for Scientific Publishing (AJSP)

In this paper precision P, recall R, and F-measure are used to calculate the performance of our algorithm, to make such measures we have to determine true positive, true negative, false positive, and false negative values for each keyword w. Tables 2 and 3 would clarify the logistic regression results with AJSP and JJSS dataset respectively.

Table 2: Logistic regression results with AJSP dataset

Documents	No. of words	No. of keywords	Accurecy	Precision	recall	F-measure	Error
Doc1	682	4	0.98	0.97	0.98	0.97	0.02
Doc2	578	5	0.98	0.82	0.90	0.85	0.02
Doc3	656	5	0.96	0.85	0.88	0.86	0.04
Doc4	651	4	0.82	0.93	0.91	0.91	0.18
Doc5	701	5	0.93	0.93	0.91	0.91	0.07
Doc6	671	9	0.98	0.92	0.92	0.92	0.02
Doc7	552	4	0.82	0.75	0.71	0.72	0.18
Doc8	561	3	0.98	0.95	0.92	0.93	0.02
Doc9	681	4	0.97	0.93	0.91	0.91	0.03
Doc10	761	6	0.88	0.80	0.88	0.83	0.12
Doc11	711	5	0.90	0.88	0.90	0.88	0.12
Doc12	552	3	0.92	0.82	0.89	0.85	0.18
Doc13	613	3	0.92	0.86	0.81	0.83	0.14
Doc14	732	4	0.92	0.91	0.82	0.86	0.09
Doc15	512	5	0.86	0.80	0.82	0.80	0.2
Doc16	703	5	0.85	0.81	0.82	0.81	0.19
Doc17	764	5	0.92	0.82	0.90	0.85	0.18
Doc18	522	5	0.94	0.91	0.89	0.89	0.09
Doc19	812	4	0.89	0.82	0.87	0.84	0.18
Doc20	643	4	0.88	0.82	0.81	0.81	0.18
Average			**0.91**	**0.86**	**0.87**	**0.86**	

Table 3: Logistic regression results with JJSS dataset

Document	No. of words	No. of keywords	Accurecy	Precision	recall	F-measure	Error
Doc1	880	4	0.97	0.80	0.91	0.85	0.03
Doc2	978	4	0.90	0.82	0.90	0.85	0.1
Doc3	856	5	0.93	0.81	0.87	0.83	0.07
Doc4	952	5	0.82	0.73	0.91	0.81	0.18
Doc5	811	5	0.92	0.83	0.91	0.86	0.08
Doc6	971	7	0.93	0.92	0.91	0.91	0.07
Doc7	692	6	0.82	0.75	0.71	0.72	0.18
Doc8	781	3	0.82	0.91	0.92	0.91	0.18
Doc9	881	4	0.93	0.91	0.90	0.90	0.07
Doc10	901	6	0.87	0.80	0.89	0.84	0.13
Doc11	1120	5	0.93	0.80	0.81	0.80	0.2
Doc12	987	5	0.92	0.85	0.89	0.86	0.15
Doc13	876	6	0.89	0.82	0.85	0.83	0.18
Doc14	655	5	0.96	0.90	0.92	0.90	0.1
Doc15	990	7	0.92	0.83	0.87	0.84	0.17
Doc16	1002	6	0.92	0.90	0.89	0.89	0.1
Doc17	876	4	0.91	0.91	0.90	0.90	0.09
Doc18	841	4	0.89	0.82	0.86	0.83	0.18
Doc19	998	4	0.91	0.80	0.81	0.80	0.2
Doc20	920	5	0.91	0.88	0.81	0.84	0.12
Average			**0.90**	**0.83**	**0.87**	**0.84**	

VII. Conclusion and Future Work

In this paper, we deal with the problem of extracting keywords from the Arabic text using a supervised learning algorithm. The performance of classification models has been evaluated using confusion matrix metrics Accuracy, precision, Recall, F-measure. We train and evaluate our model with the Logistic Regression average accuracy is found to be 0.91 with precision 0.86, recall 0.87 and f-measure 0.86 for the AJSP dataset and average accuracy found to be 0.90 with precision 0.83, recall 0.87, and f-measure 0.84 for the JJSS dataset. Also, these are promising results in the field of extracting keywords from the Arabic text. In the meantime, we will put on this method on some standard documents corpus. It will also be stimulating to apply the AKE model to a great number of text mining applications, such as text classification, clustering, summarization, and filtering. We recommend that the use of the same features be tested on other document datasets, and see if results, especially in extracting rules, might be drastically different. We recommend the use of more features to improve the keyword extraction process.

Conflict of Interest:

There was no relevant conflict of interest in this paper

References

I. Aarti Sangwan, Partha Pratim Bhattacharya, “A Hybrid Cryptography and Authentication based Security Model for Clustered WBAN”, J.Mech.Cont.& Math. Sci., Vol.-13, No.-1, March – April (2018) Pages 34-54

II. A. A. Awajan, "Unsupervised Approach for Automatic Keyword Extraction from Arabic Documents". In Proceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014), pp. 175-184.2014.

III. A.Bilski, "A review of artificial intelligence algorithms in document classification". International Journal of Electronics and Telecommunications, Vol. 57, Issue 3, pp. 263-270, 2011.

IV. B. Armouty, and S.Tedmori, "Automated Keyword Extraction using Support Vector Machine from Arabic News Documents". In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), pp. 342-346. IEEE, 2019.

V. D. Suleiman, and A. Awajan, "Bag-of-concept based keyword extraction from Arabic documents". In 2017 8th International Conference on Information Technology (ICIT), pp. 863-869, 2017.

VI. C. Zhang, "Automatic keyword extraction from documents using conditional random fields". Journal of Computational Information Systems, Vol. 4, Issue 3, pp. 1169-1180, 2008.

VII. D. Suleiman, and A. Awajan, "Bag-of-concept based keyword extraction from Arabic documents". In 2017 8th International Conference on Information Technology (ICIT), pp. 863-869, 2017.

VIII. E.H. Omoush, and V.W. Samawi, "Arabic keyword extraction using SOM neural network". International Journal of Advanced Studies in Computers, Science and Engineering, Vol. 5, Issue 11, pp. 7, 2016.

IX. F. Sebastiani, "Machine learning in automated text categorization". ACM computing surveys (CSUR), 2002. Vol. 34. Issue 1, p. 1-47, 2002.

X. K. Sarkar, M. Nasipuri, and S. Ghose "A new approach to keyphrase extraction using neural networks". arXiv preprint arXiv:1004.3274, 2010.

XI. Kesana Mohana Lakshmi, Tummala Ranga Babu, “Robust Algorithm for Telugu Word Image Retrieval and Recognition”, Robust Algorithm for Telugu Word Image Retrieval and Recognition, J.Mech.Cont.& Math. Sci., Vol.-14, No.-1, January-February (2019) pp 220-240

XI. M. Al-Kabi, H. Al-Belaili, B. Abul-Huda, and A. H. Wahbeh, " Keyword extraction based on word co-occurrence statistical information for Arabic text". Abhath Al-Yarmouk" Basic Sci. Eng, Vol. 22, Issue 1, pp: 75-95,2013.

XII. M. M. Abdulwahid, O. A. S. Al-Ani, M. F. Mosleh, and R. A. Abd-Alhmeed. "Optimal access point location algorithm based real measurement for indoor communication". In Proceedings of the International Conference on Information and Communication Technology, pp: 49-55, 2019.

XIII. M. Labidi, "New Combined Method to Improve Arabic POS Tagging". Journal of Autonomous Intelligence, Vol. 1, Issue 2, pp.23-28, 2019.

XIV. P.-I. Chen, and S.-J. Lin, "Automatic keyword prediction using Google similarity distance". Expert Systems with Applications. Vol. 37, issue 3, pp. 1928-1938, 2010.

XV. R. Feldman, and J. Sanger, "The text mining handbook: advanced approaches in analyzing unstructured data". 2007: Cambridge university press.

XVI. R.M. Alguliev, and R.M. Aliguliyev. "Effective summarization method of text documents". The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). 2005.

XVII. Sallam, R.M., H.M. Mousa, and M. Hussein, "Improving Arabic text categorization using normalization and stemming techniques". International Journal of Computer Applications, Vol. 135, Issue 2, pp. 38-43, 2016.

XVIII. S. K. Shevade and S. S. Keerthi. "A simple and efficient algorithm for gene selection using sparse logistic regression". Bioinformatics, Vol. 19, Issue 17, pp: 2246-2253.

XIX. S. Lee, I., Lee, H., Abbeel, P., and A. Y. Ng, "Efficient l~ 1 regularized logistic regression. In AAAI", Vol. 6, pp. 401-408, 2016.

XX. T. Jo, "Neural based approach to keyword extraction from documents ". In International Conference on Computational Science and Its Applications. 2003. Springer.

XXI. V. Singh B. Kumar, and T. Patnaik, "Feature extraction techniques for handwritten text in various scripts: a survey". International Journal of Soft Computing and Engineering (IJSCE), Vol. 3, Issue 1: pp. 238-241, 2013.

XXII. محقق, ن., نیلوفر, اطلسی, علی بیک, صالحی, حجتی زاده, ... & باقری . "The Relationship between Number of Keywords Used in Titles of Articles and Number of Citations to These Articles in Selected Journals Published by Tehran University of Medical Sciences". مطالعات کتابداری و علم اطلاعات,

XXIII. Y. Wang and X.-J. Wang. "A new approach to feature selection in text classification". in 2005 International conference on machine learning and cybernetics. 2005.

XXIV. Y. Ying, Qingping, T. Qinzheng, Z. Ping, and L. Panpan "A graph-based approach of automatic keyphrase extraction". Procedia Computer Science, Vol. 107, pp. 248-255, 2017.