DOI: 10.24996/ijs.2022.63.1.28

ISSN: 0067-2904

Arabic Keywords Extraction using Conventional Neural Network

Nada A.Z. Abdullah Noor T. Jaboory

Department of Computer Sciences, College of Science, University of Baghdad, Baghdad, Iraq

Received: 19/9/2020 Accepted: 10/11/2020

Abstract

Keywords provide the reader with a summary of the contents of the document and play a significant role in information retrieval systems, especially in search engine optimization and bibliographic databases. Furthermore keywords help to classify the document into the related topic. Keywords extraction included manual extracting depends on the content of the document or article and the judgment of its author. Manual extracting of keywords is costly, consumes effort and time, and error probability. In this research an automatic Arabic keywords extraction model based on deep learning algorithms is proposed. The model consists of three main steps: preprocessing, feature extraction and classification to classify the document tokens into keyword or not, Conventional Neural Networks (CNN) is used as a classifier.

Two types of dataset are building in this research to test the proposed model, the first dataset form Arab Journal for Scientific Publishing (AJSP), the other dataset from Jordan Journal of Social Sciences (JJSS). The experiment results indicate promising results in the field of Arabic keyword extraction; the average accuracy of Conventional Neural Networks is found 0.97 with average precision 0.92.

Keywords: Arabic Keyword Extraction, deep learning, CNN

أستخلاص الكلمات الدلالية العربية باعتماد الشبكات العصبية الالتفافية

ندا عبد الزهرة ، نور جبوري

قسم علوم الحاسبات، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

تزود الكلمات المفتاحية القارئ بملخص لمحتويات الوثيقة وتلعب دورًا هامًا في أنظمة استرجاع المعلومات ، خاصة في تحسين محركات البحث وقواعد البيانات الببليوغرافية. علاوة على ذلك ، تساعد الكلمات الرئيسية في تصنيف المستند إلى موضوع ذي صلة. استخراج الكلمات المفتاحية يدويا" يعتمد على محتوى المستند أو المقال وحكم مؤلفه بالاضافة الى كونه مكلف ، ويستهلك الجهد والوقت ، ومن المحتمل حدوث خطأ. في هذا البحث ، يُقترح نموذج استخراج تلقائي للكلمات الرئيسية العربية يعتمد على خوارزميات التعلم العميق. يتكون النموذج من ثلاث خطوات رئيسية: المعالجة المسبقة ، واستخراج الميزات ، والتصنيف لتصنيف الرموز المميزة للمستند إلى كلمات رئيسية أم لا ، ويتم استخدام الشبكات العصبية التقليدية. (CNN) كمصنف.

تم بناء مجموعتين من البيانات في هذا البحث لاختبار النموذج المقترح ، المجموعة الأولى من المجلة العربية للنشر العلمي (AJSP) ، ومجموعة البيانات الأخرى من المجلة الأردنية للعلوم الاجتماعية(JJSS) .

*Email: Nada.Abdullah@sc.uobaghdad.edu.iq

أظهرت نتائج التجربة نتائج واعدة في مجال استخلاص الكلمات المفتاحية العربية حيث تم تحقيق متوسط صحة 0.97 ومتوسط دقة 0.92.

1. Introduction

A large number of documents are existed in the current time; as it is available in an electronic form. The fast growth of the internet use with its unstructured contents leading to increase the keyword extraction importance, keyword gives a clue to the reader about the document contents [1]. Not all of the words are reflecting the contents of the document; it is required for selecting significant words that connect with its contents the selected words are called keywords [2]. The access to large amount of documents that doesn’t have keywords and available online is limited [3].

The automatically occurrence of keyword is considered to be a very important task in text mining by replacing the manual keywords extraction and it is considered as a process of time consuming. The English documents keyword extraction is not a new subject; it is implemented by using different techniques as reported in [4].

The proposed in this paper is a model of Arabic keyword extraction from Arabic document; this model contains of three stages: the first stage is preprocessing to sift text from Punctuation marks, Symbols and numbers. The second stage is features extraction, in this work (eight) features are used; first is Term Frequency (TF) procedure which weights the words, others are related to the position of the word in title, abstract and the way in which the word is printed. The third stage is a classification of words using CNN algorithm, which is classified as a one of deep learning algorithms.

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

2. Related Work

This section includes a number of recent studies that proposed for keyword extraction for Arabic language documents; the researches in this field are few. The following are some of these researches:

In 2013, Al-Kabi, et al. presented a study about a particular system of keyword extraction for Arabic documents using co-occurrence statistical information used in English and Chinese language systems. The basic of the proposed method is to extract top frequent terms and build a matrix of co-occurrence. If the co-occurrence for a particular term is in the degree of biasness, that term is an important term and it is mostly a keyword. Other different novel methods use frequency-inverted term and term frequency (TF-ITF). The obtained results of this method is better than the method of TF-ITF, the accuracy was 0.58 and the recall was 0.63, while second experiment achieved 64% of accuracy. Two experiments results showed ability of this method to be applied for Arabic documents and its performance is acceptable compared to other techniques [3].

In 2014, Awjan, A.A. made a study that used unsupervised approach consists of two-phases for keyword extraction from a document written in Arabic that contains statistical analysis and linguistic information. First phase is detecting all the N grams that could be taken as keywords. Second phase is analyses of the N grams by use of morphological analyzer for replacing the N grams words with their basic forms that considered derived words roots and body of nob-derivative words. The N grams that contain a similar base form are regrouping with an accumulation for their counts. The proposed work results are achieved 0.51 of accuracy. The experiments of the proposed work are extracting 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 [4].

In 2016, Omoush, E.H. and Samawi V.W presented a study using self-organization map (SOM) neural network as a method of unsupervised learning. The proposed method performance used F-measure, recall and a precision for the evaluation. This technique used two datasets. The first is JJSS dataset and the second is Wikipedia dataset. The obtained results showed a precision of 42.84% by using JJSS dataset and 46% by using the Wikipedia dataset [5].

In 2017, Suleiman, D. and Awajan. A.A presented a study about a new method of keyword extraction using the bag-of-concept for extracting keywords from Arabic text. Algorithm proposed in that study was utilizing the semantic vector space model instead of traditional vector space model to group words into classes. The word context matrix is built by the new method and the synonyms words were grouped in the same class. The new approach evaluations lead to the use of dataset that contains three documents and can be compared with keywords extraction using method of equivalence classes' term from Arabic documents. The obtained results from the proposed method had 90% of precision in the second document and it is considered the best because the keywords number is small [6].

In 2019 Armouty, B., and Tedmori, S. made a study about keyword extraction from Arabic document. The proposed method used statistical features, supervised learning technique and Support Vector Machine classifier. This method applied on Arabic news documents, the result showed a precision of 0.77 and a recall of 0.58 [7].

2. The principle Challenges of Arabic Language

There are many challenges in Arabic language related to keyword extraction. The following are main challenges: [8]

- There are specific combinations of Arabic characters combinations that are not unique in rendition. For example, symbols that combined name HAMZA with the "أ" name HAMZA dropping the "ا".
- The Arabic language has high inflection degree. As example gives the possessive, word should have letter "ي" called YA; it is connected to the end of Arabic word. The disjoint does not exist in Arabic language such as "MY".
- The English language has broken plurals with resemblance to a specific form that singular. Broken plurals in Arabic language are not the same as the English language. The broken plurals in Arabic language are not bounded to the morphological rules, and also they are hard to be related to singular form.
- The words in Arabic language are often derived from a simple bare the verbs called the roots that containing three letters. One letter from the roots could be dropped or sometimes more than one letter in some derivations. Sometimes the root tracing that derived from a particular word can be a big problem.
- The vowels often deleted from written Arabic words. This deletion of that vowels leads to ambiguity through the interpretation of the words.
- The synonyms is widely known and used in the all kind of Arabic literature. To make a better recall, the synonyms must be considered during the processing of query.

4. The Proposed Automatic Keyword Extraction Model (AKE)

The model of AKE consists of three phases: first phase is the preprocessing phase; it implements tokenization, performing the normalization and filtering tokens from stop word and stemming. Second phase is feature extraction phase, and final phase is decision making phase. In the decision-making phase every word in word-list will be classified either as a keyword or not a keyword. The proposed solution uses CNN as a classifier. Figure 1 illustrates the proposed main block diagram.

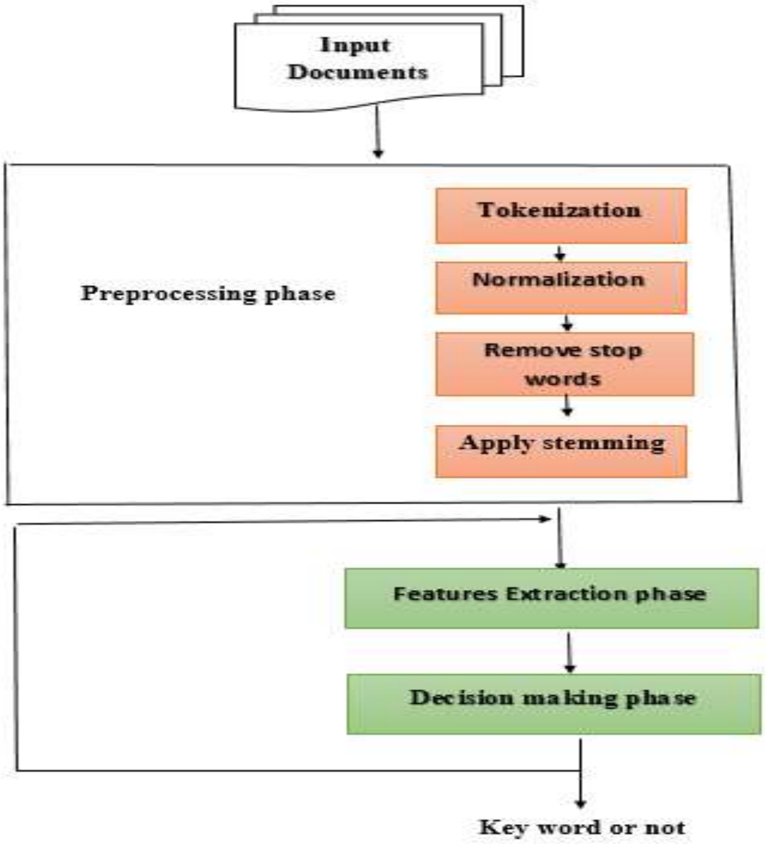

Figure 1-General block diagram of the proposed model

4.1 Preprocessing Phase

The preprocessing is used to remove factors of language-dependent from text, and consists of a removal of stop words, tokenization, and stemming [9]. The preprocessing of document or dimensionality reduction (DR) allows skilled data manipulation for text categorization. The dimensionality reduction considers an essential step during the process of classification. It allows deletion of unimportant features of document, hence, that lead sometimes to minimize the efficiency of the classification, also minimize their accuracy and speed [10]. In the stage of preprocessing, text is split into tokens. This operation is called tokenization to create 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 [11]. The preprocessing used in this work illustrated in Figure 2.

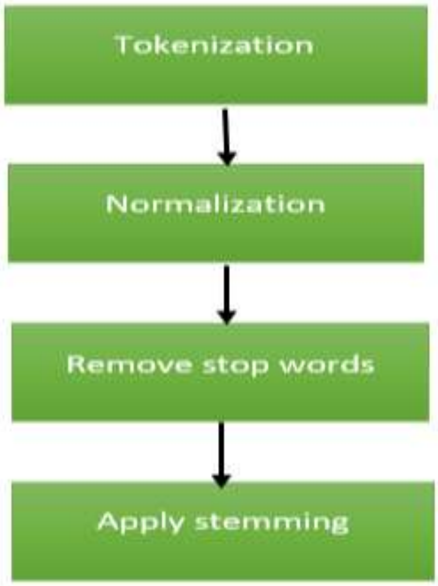

Figure 2- Preprocessing Steps

Algorithm 1 illustrates the preprocessing basic steps. The algorithm is 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 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 stop words
     Step 5: Stemming the residual token// extract the root of the tokens by using ISRI stemming algorithm
Return (list of tokens )
End
```

4. 2 Features Extraction Phase

In feature extraction phase (eight) properties are used as features for every token. These features are summarized in the Table 1.

Table 1- features extraction

No	Features	Descriptions	Regularization Method
1	TF	The part of the number of epochs a token occurs in the text to total of tokens in the documents.	$TF_{i,j} = \frac{n_{i,j}}{\Sigma_k n_{i,j}}$
2	T	If the token has showed in the title	[0,1]
3	A	If the token has showed in the abstract	[0,1]
4	Fs	If the word has showed in the first sentence	[0,1]
5	B	If the word is bold	[0,1]
6	*I*	If the word is italic	[0,1]
7	U	If the word is under line	[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]

5. Convolutional Neural Networks

In this work, CNN is used to classify text tokens as keyword or not. CNN was inspired by the mechanism of Biological natural visual cognitive recognition. In 1959, Hubel and Wiesel found the mechanism behind information processing of the visual system. They found out that the visual cortex for animals is constructed with layers and neurons. The neuron only reacts to a receptive field. Different neurons are responsible for different receptive fields and the final vision of an animal is the combination of these fields. Enlightened by this idea, Le Cun et al. in 1998, designed a multilayer Neural Networks to classify manuscript numbers, which became the modern structure of CNN that is widely used today. CNN is a leading architecture in extracting functional features. As a matter of fact, the inputs of NLP tasks are either sentences or documents that are broken by tokens during the data pre-processing stage. Each token stands for each word; in order to process these words with a computer, the tokens translate into a numeric format, which is similar to the representation of images. The input of NLP tasks is constructed into a word vector. [10]

A CNN involves an input and an output layer, with multiple hidden layers. The hidden layers characteristically contain a sequence of convolutional layers. The activation function is mostly a RELU layer, and then tracked by other convolutions such as pooling layers, wholly connected layers and normalization layers, mentioned as hidden layers since their inputs and outputs are disguised by the activation function and final convolution. The architecture model is shown in Figure 3.

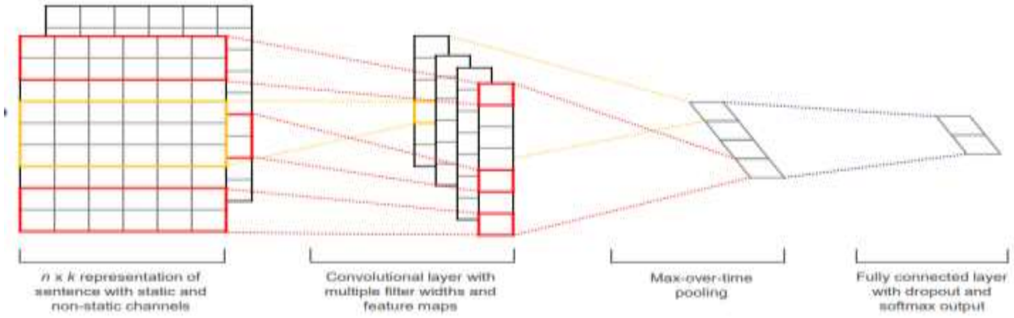
n x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

Figure 3- A simple example of CNN architecture for classifying texts

5.1 CNN for text classification

In this work, CNN is used as classifier for Arabic text tokens as illustrated in Figure 4. It starts with an input sentence broken up into words or word embedding .Words are broken up into features and fed into a convolutional layer. The results of the convolution are “pooled” or aggregated to a representative number. This number is fed to a fully connected neural structure, which makes a classification decision based on the weights assigned to each feature within the text.

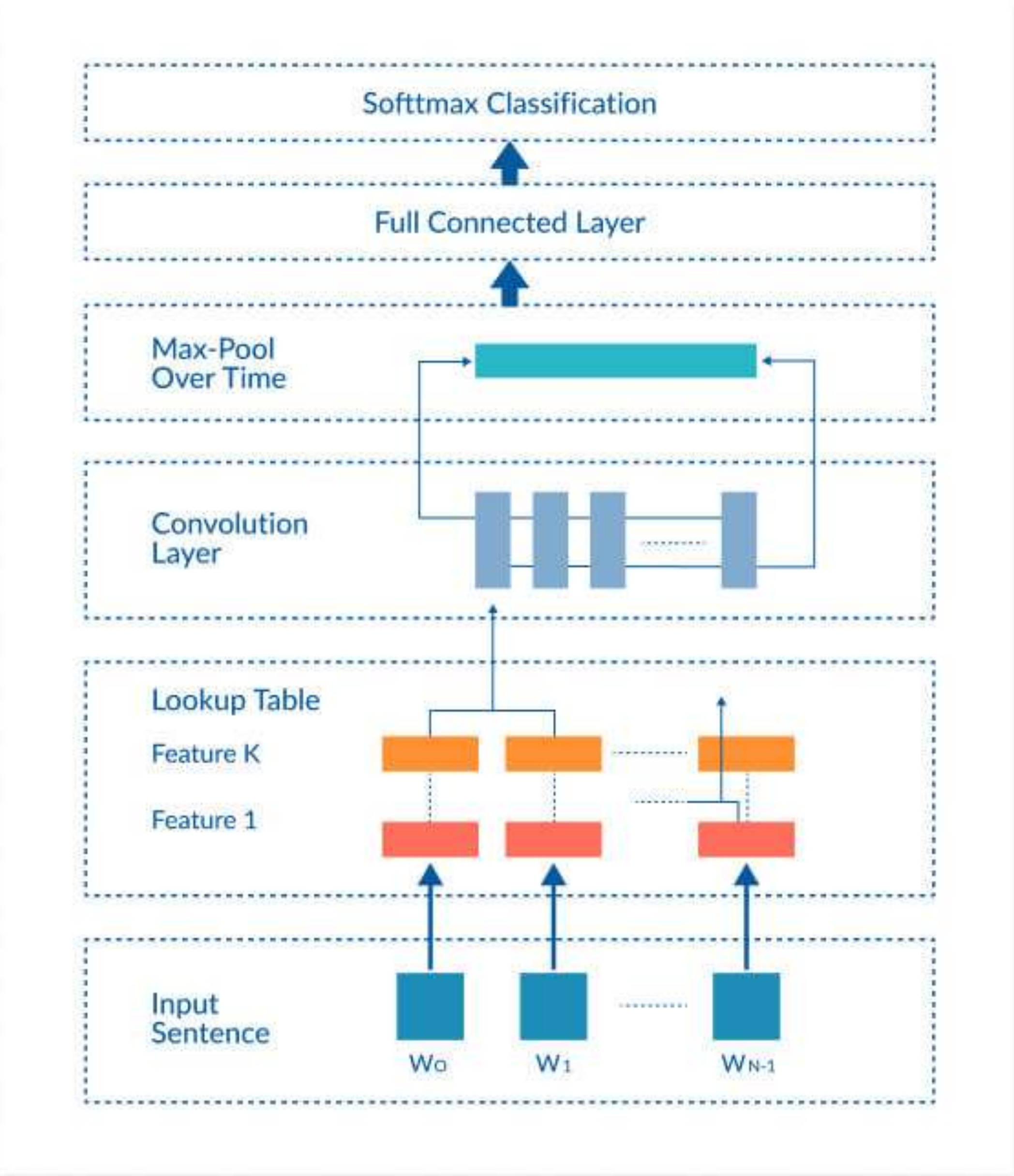

Figure 4- CNN architecture for text classification

Embedded layer is initialized with random weights and will learn for all of words in the training dataset. It is a flexible layer that can be used in a variety of ways, such as:

- It can be used alone to learn a word embedding that can be saved and used in another model later.
- It can be used as part of a deep learning model where the embedding is learned along with the model itself.
- It can be used to load a pre-trained word embedding model, a type of transfer learning.

The Embedding layer is defined as the first hidden layer of a network. It must specify the following:

- **input_dimension** : This is the size of vocabulary in the text data. For example, if your data is integer encoded to values between 0-10, then size of the vocabulary would be 11 words.
- **output_dimension**: This is the size of the vector space in which words will be embedded. It defines the size of the output vectors from this layer for each word. For example, it could be 32 or 100 or even larger. Test different values for your problem.
- **input_length**: This is the length of input sequences, as you would define for any input layer of a Keras model. For example, if all of your input documents are comprised of 1000 words, this would be 1000.

Keras provides a set of different convolutional layers that can be used. In case of working with text, the convolutional layer that is used for this task is conv1D. This layer has several parameters. The important ones are the number of filters, the kernel size, and the activation function. This layer can be added between the Embedding layer and the GlobalMaxPool1D layer. [13].

Figure 5 Illustrates how this convolution works which starts by taking a spot of input features with the size of the filtered core. Using this spot, the dot product of the multiplied weights of the filter is taken. [13].

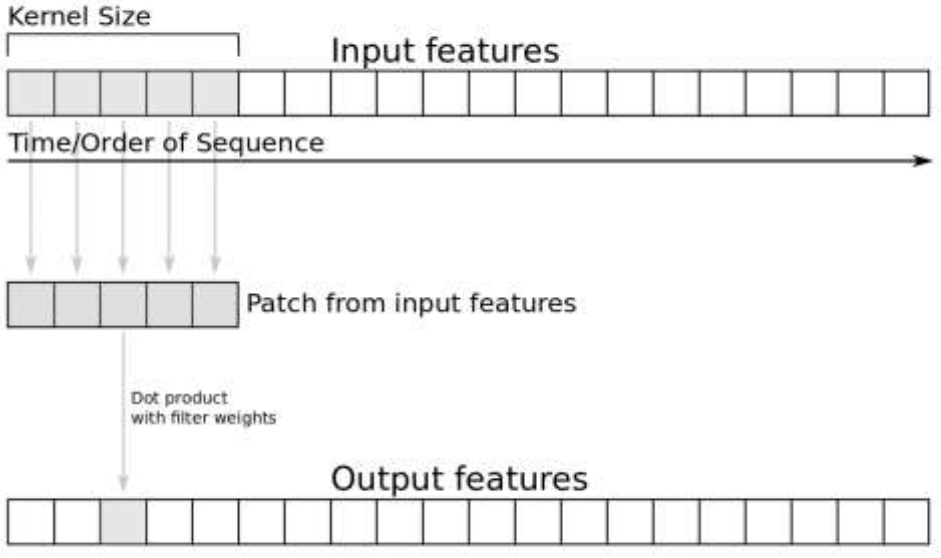

Figure 5-1D convolution

Algorithm 2 illustrates the steps to implement Arabic keyword extraction model using CNN.

Algorithm (2) CNN algorithm for Keyword Extraction

```
Input: Features of the document tokens
Output: classification keyword or not keyword
Begin
Step1: read features file

Step2: separate the input features and the target variable
Step3: split the train and test validation set
Step4: define the model
 -Initializing the Sequential nature for CNN model
  -Adding the embedding layer
  Adding the convolutional layer 1D
  add(layers.GlobalMaxPooling1D())
  model.add(layers.Dense(10, activation='relu'))
  model.add(layers.Dense(1, activation='sigmoid'))
Step 5: compile the model
binary_crossentropy(negative log-Loss)
The Optimizer using 'Adam' (Adaptive Moment Estimation)
Metrics', which is used to judge the performance of our model.
Matrix [accuracy]
Step6: train the model
    Fit method the training set (x_train,y_train) for training the model

 Run the model for 500 epochs
  Print accuracy
End
```

6. **Model Evaluation**

Two types of dataset were used, the Arab Journal for Scientific Publishing (AJSP), and the other dataset from Jordan Journal of Social Sciences (JJSS). 20 academic papers were taken from each journal. We arbitrarily chose academic papers for every text comprises the title, abstract, keywords full-document, borders information of sections, references. These texts have rich semantics topographies and appropriate to achieve classification of keywords in good form. Table 2 shows accuracy, precision score, recall score, F-measure and error for every document in the AJSP dataset.

Table 2- CNN result with AJSP

Documents	No. of words	No. of keywords	Accurecy	Precision	Recall	F- measure	Error
Doc1	577	4	0.99	0.98	0.97	0.97	0.01
Doc2	578	5	0.98	0.97	0.97	0.97	0.02
Doc3	656	5	0.98	0.85	0.89	0.86	0.02
Doc4	651	4	0.97	0.89	0.90	0.89	0.03
Doc5	701	5	0.98	0.95	0.93	0.93	0.02
Doc6	671	8	0.99	0.96	0.93	0.94	0.01
Doc7	552	4	0.89	0.91	0.90	0.90	0.11
Doc8	561	4	0.98	0.93	0.92	0.92	0.02
Doc9	681	4	0.98	0.91	0.90	0.90	0.02
Doc10	761	6	0.96	0.89	0.90	0.89	0.04
Doc11	711	4	0.97	0.90	0.88	0.88	0.03
Doc12	552	4	0.96	0.92	0.94	0.92	0.04
Doc13	613	3	0.93	0.93	0.93	0.93	0.07
Doc14	732	4	0.94	0.88	0.92	0.89	0.06
Doc15	512	4	0.93	0.95	0.90	0.92	0.07
Doc16	703	5	0.94	0.89	0.88	0.88	0.06

Doc17	764	5	0.95	0.96	0.91	0.93	0.05
Doc18	522	5	0.97	0.91	0.97	0.93	0.03
Doc19	812	4	0.93	0.89	0.88	0.88	0.07
Doc20	643	4	0.94	0.88	0.91	0.89	0.06
Average			**0.95**	**0.91**	**0.91**	**0.91**	

Table 3 Shows accuracy, precision score, recall score, F-measure and error for every document in the JJSS dataset.

Table 3- CNN result with JJSS

Documents	No. of words	No. of keywords	Accurecy	Precision	Recall	F- measure	Error
Doc1	880	4	0.97	0.92	0.91	0.91	0.03
Doc2	978	5	0.95	0.91	0.93	0.91	0.05
Doc3	856	5	0.96	0.86	0.95	0.90	0.04
Doc4	952	4	0.92	0.88	0.91	0.89	0.08
Doc5	811	5	0.90	0.90	0.91	0.90	0.1
Doc6	971	9	0.93	0.93	0.91	0.91	0.1
Doc7	692	4	0.89	0.89	0.89	0.89	0.11
Doc8	781	3	0.94	0.91	0.88	0.89	0.06
Doc9	881	4	0.97	0.92	0.88	0.89	0.03
Doc10	901	6	0.97	0.93	0.91	0.91	0.03
Doc11	1120	5	0.94	0.92	0.80	0.85	0.06
Doc12	987	5	0.93	0.93	0.82	0.87	0.07
Doc13	876	6	0.92	0.93	0.94	0.93	0.08
Doc14	655	5	0.88	0.94	0.85	0.89	0.12
Doc15	990	7	0.89	0.92	0.87	0.89	0.11
Doc16	1002	6	0.95	0.89	0.89	0.89	0.05
Doc17	876	4	0.96	0.89	0.93	0.90	0.04
Doc18	841	4	0.97	0.94	0.95	0.94	0.03
Doc19	998	4	0.89	0.89	0.90	0.89	0.11
Doc20	920	5	0.97	0.94	0.89	0.91	0.08
Average			**0.93**	**0.91**	**0.89**	**0.89**	

7. Conclusion and Recommendations

In this paper, we deal with the problem of extracting keywords from an Arabic text using one of the deep learning algorithms CNN. Performance of the proposed model has been evaluated using accuracy, precision, recall, and F-measure. We train and evaluate our model with the CNN deep learning algorithm which gave best results. This is an indication that deep learning is successful in classification problems. It is known that CCN is mainly used to classify image data. However, its use in classifying text has shown great results on text data as well.

A recommendation for future work, use the same set of features and test it on another dataset then comparing the results. To improve the performance of keyword extraction system, more features can be used.

Documents used were in PDF format, which required converting it to TXT format. This caused some noise therefore, it is better to use HTML crawl content to produce a less noisy dataset. In addition to saving effort to reduce the noise generated by converting the PDF to TXT format. While proposed CNN model uses one convolutional layer 1D and one max-pooling layer, it would be stimulating to explore the uses of more layers in future work.

References

[1] Sarkar, K., M. Nasipuri, and S. Ghose, "A New Approach To Keyphrase Extraction Using Neural Networks", *arXiv preprint arXiv*:1004.3274, 2010.

[2] Jo, T. ,"Neural Based Approach to Keyword Extraction from Documents", International Conference on Computational Science and Its Applications. 2003. Springer.

[3] Al-Kabi M., Al-Belaili H., Abul-Huda H. and Wahbeh A. H.," Keyword Extraction Based nn Word Co-Occurrence Statistical Information For Arabic Text", *Abhath Al-Yarmouk Basic Sci. Eng*, vol. 22, no. 1, pp. 75-95, 2013.

[4] Awajan, A.A., "Unsupervised Approach for Automatic Keyword Extraction from Arabic Documents",Proceedings of the 26th Conference on Computational Linguistics and Speech Processing (ROCLING 2014). pp. 175-184, 2014.

[5] Omoush, E.H. and Samawi V.W., "Arabic Keyword Extraction Using SOM Neural Network", *International Journal of Advanced Studies in Computers, Science and Engineering*, vol. 5, no. 11, pp. 7, 2016.

[6] Suleiman, D. and Awajan. A.A., "Bag-of-Concept Based Keyword Extraction from Arabic Documents",2017 8th International Conference on Information Technology (ICIT). 2017. IEEE.

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

[8] Feldman, R. and J. Sanger, "*The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data*", *Cambridge University Press*, 2007.

[9] Wang, Y. and X.-J. Wang., "A New Approach To Feature Selection In Text Classification",2005 International Conference On Machine Learning And Cybernetics. 2005. IEEE.

[10] Bilski, A., "A Review Of Artificial Intelligence Algorithms In Document Classification", *International Journal of Electronics and Telecommunications*, vol. 57, no. 3, pp. 263-270, 2011.

[11] Ramasubramanian, C. and R. Ramya, "Effective Pre-Processing Activities In Text Mining Using Improved Porter's Stemming Algorithm", *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 12, pp. 4536-4538, 2013.

[12] 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, no. 2, pp. 38-43, 2016.

[13] Singh, V., 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, no. 1, pp. 238-241, 2013.

[14] WANG, Jingjing, "Using Convolutional Neural Networks to Extract Keywords and Key Phrases About Foodborne Illnesses", 2019.