Journal of Theoretical and Applied Information Technology
30th June 2018. Vol.96. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTOMATIC MACHINE LEARNING TECHNIQUES (AMLT) FOR ARABIC TEXT CLASSIFICATION BASED ON TERM COLLOCATIONS

[1]FEKRY OLAYAH, [2]WASEEM ALROMIMA
[1]Faculty of Computers Science & Information System, Najran University
Najran, Saudi Arabia
[2]Department of Computers Science and Information, Taibah University
Madinah, Saudi Arabia
[1]Dr.Fekry_Olayah@yahoo.com, [1]faahmed@nu.edu.sa, [2]waseem.2020@yahoo.com, [2]wromema@taibahu.edu.sa

ABSTRACT

Due to the rapid and increased availability of documents in a digital format, effect for retrieving information with highest accuracy and the lowest error rate is becoming more difficult. Text Classification (TC) has become one of the key techniques for controlling and organizing documents based on the content of documents. Therefore, keyword extraction is one of the most important natural language processing applications, which extracts information from the document such as term collocations, which are two or more words appear together and always seem as associated. In Arabic language, there are many problems in keyword extraction because of the complexity of Arabic orthography. Moreover, the accuracy is affecting by the document content and the classification technique used. The need for automatic text classification came from a large amount of electronic documents on the web. This research aims to propose an Automatic Machine Learning Techniques (AMLT) for classifying Arabic documents by using term collocations. These collocations are mined from Arabic documents, the extracted term collocations will scoring by using association measure and will be used as terms feature selection. To achieve this study, we used Arabic documents divided into four categories (Economy/ business, Politics, Religion and Science). The results of our approach have compared with the full-document approach and summary-document approach using four techniques (SVM, NB, J48, and KNN) for Arabic documents to determine which classifier is more accurate for Arabic text based on term collocation. The evaluation results proved that our proposed approach outperforms the other method in accuracy.

Keywords: *Arabic Language, Text classification, Term collocations, bi-gram, Machine Learning, Category*

1. INTRODUCTION

Accessibility of information in the web is the main feature in knowledge acquisition. The World Wide Web (WWW) has become a vast library of unstructured data, which is laboriously understanding and processed without using intelligent techniques. Although search engines (SEs) are main tools for retrieving information published on the Web [4], existing SEs still have various significant limitations. It is extremely strenuous to help people locate the needed information. Therefore, intelligent methods for text classification are needed to enhance the current SEs, especially for Arabic language text because the complexity orthography of Arabic language.

The volume of Arabic content is growing, there are more than 100 million Arabic web pages covering various topics, such as business, science, politics, and religion [5, 6]. In addition, the Arabic language is one of the Semitic languages, and more than 422 million people [7] speak it.

Arabic Natural Language Processing (ANLP) still in its open research compared to the works that have done in the Latin language, because there are some issues that slow down progress in ANLP mentioned in [10]. These issues illustrate as the following:

- Arabic is highly inflectional and derivational, which makes morphological analysis a very complex task.
- The absence of diacritics in the written text creates ambiguity and, therefore, complex

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

morphological rules are required in order to identify the tokens and parse the text.
- Diacritics are not consistent or predictably marked with Arabic letters.
- The writing direction is from right-to-left and some of the characters change their shapes depending on their location in the word.
- Capitalization is not use in the Arabic language, which makes it hard to identify proper nouns and abbreviations.

On the other hand, text classification or text categorization task is assigning a document to one or more predefined classes or categories. It has gained an important in the researches in wide range of applications, which include text summarization, Proper Noun Extraction, Named Entity, etc. [11]. Moreover, Information Extraction is one of NLP application, which extracts term collocations and information from the dataset. Term Collocations is define as a sequence of terms co-occurs together in the corpus [12]. In our approach, we used the term collocations in the segmentation of the text into terms based on N-gram model, such as the bi-gram, which has two levels of gram.

Shallow neural networks and other machine learning algorithms have been introduce in order to solve the Arabic text classification problem. Recently, further improvement has been introduce using more complex models that carry on account Arabic text characteristics.

This paper introduces a new approach for text classification to improve classify Arabic text results. The model uses the n-gram term collocations that appear as a sequence of terms in the dataset (category). The n-gram term collocations are extracts from the Arabic dataset automatically. We applied the proposed approach to the Arabic documents, which contains four categories (economy, politics, religion, and science). The Vector Space Model (VSM) has employed to represent both documents and terms features. The proposed method does not need any external resource. Therefore, the difficulty and complexity of the preprocessing process for an ambiguous language like Arabic. The results of the experiments indicate that this approach performs quite well when compared to traditional approaches and summarization approach.

This paper is orderly as follows. Section 2 presents the background and related work, where section 3 presents the proposed approach methodology. Section 4 describes the implementation of the proposed approach. Section 5 discusses the experiments and evaluation for the proposed. Finally, section 6 concludes the paper.

2. RELATED WORK

Many approaches have executed to explain the problem of text classification. There are a scarce researches tackled the area of Arabic text classification because of the complexity of the Arabic grammar. According to [3], the authors used statistical classifier based on Nave Bayesian (NB) to categorize Arabic web documents; the approach used the root stemmer to extract the roots of the words.

Al-Shalabi and Obeidat [8] implemented a text classification system for Arabic. This system compares the representation of document by N-grams (unigrams and bigrams) and single terms (bag of words) as a feature extraction method in the pre-processing step. Afterwards, he used Term Frequency and Inverse Document Frequency (TFIDF) to reduce dimensionality and K-Nearest Neighbors (KNN) as classifier for Arabic text classification. The experimental results showed that using unigrams and bigrams as representation of documents outperformed the use of bag of words in terms of accuracy. The work presented in the [12] provides how to extracts n-gram terms collocations based on tagging sequences (from 2- 6 gram) from Arabic Quran corpus. Authors proposed a prototype that extracts collocation by matching the input structured pattern of Arabic language versus the Part of Speech Tagging. This approach is useful for the automatic extended query, grammatical relations, and construction of domain ontology.

In paper [13], the authors presented a hybrid approach to extract Multi-Word Terms (MWT) from Arabic corpus. The authors extracted compound nouns as an important type of MWT and selected a bi-gram terms. Two main filtering used, starting from linguistic filter (preprocessing, extract of noun sequence as well as noun connected by prepositional, and extract of bi-gram) and ending with statistical filter: rank the bi-gram based on Log-Likelihood Ratio (LLR) and C-value method. In [14], a statistical method called maximum entropy is using to classify Arabic News article. In [16], the authors designed a multi-word term extraction program for Arabic language. They used a hybrid method to extract multi-word terminology from Arabic corpus. They used some linguistic information to extract and filter the candidates of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

multiword terminology. In the paper [17], the authors used Vector Space Model (VSM) to classify the Arabic text and compared result with K-Nearest Neighbor classifier.

Moreover, Arabic Topic Identification Algorithm involves the identification and selection of topics from an Arabic document. It is achieve through various tools including the VSM. Single document summarizer performed using the Vector Space Model, which takes Arabic document and the initial sentence of the file to generate an executive summary. The automatic text summarization emerges as an important tool in the Arabic Topic Detection. According to the study conducted by Koulali, El-Haj et al. [18], the use of this model helped in the delivery of results that were comparable to the results achieved through the application of the full-text manuscript collection. The automatic text summarization is an effective method for the selection of characters and reduces noise information in the Arabic document [18].

On the other hand, topic identification can completed using the K-Nearest Neighbor classifier, which based on the Tf-idf cloud compute architecture. Each topic is name by the N documents using a training process [19]. The Tf-Idf algorithm allows the computation of the word weights of the document in which the training documents is present as a vector weights.

To sum up, many of researchers have conducted great research to upgrade text classification for Arabic text, although more research efforts still needed because of the complexity of the Arabic language and the lack of resources. In this paper, we present new approach for Arabic text classification, which depends on term collocation extracted from the Arabic dataset.

3. THE PROPOSED APPROACH METHODOLOGY

The architecture of the proposed approach Arabic text classification is show in Figure 1. It shows the methodology that we have followed to enhance the Arabic text classification process. The Methodology contains five levels: the first, documents preprocessing, which tokenize the documents that has created and maintained for the Arabic dataset categories. The second stage is n-gram term collocation extraction, where collocations term are extracting from the Arabic dataset categories. Third, feature selection, which eliminates large number of terms based on the associated computation. Four stage is machine-learning classifier, which the classification algorithms are use such as SVM, 48J, etc. Details for each stags are discussion as follows.

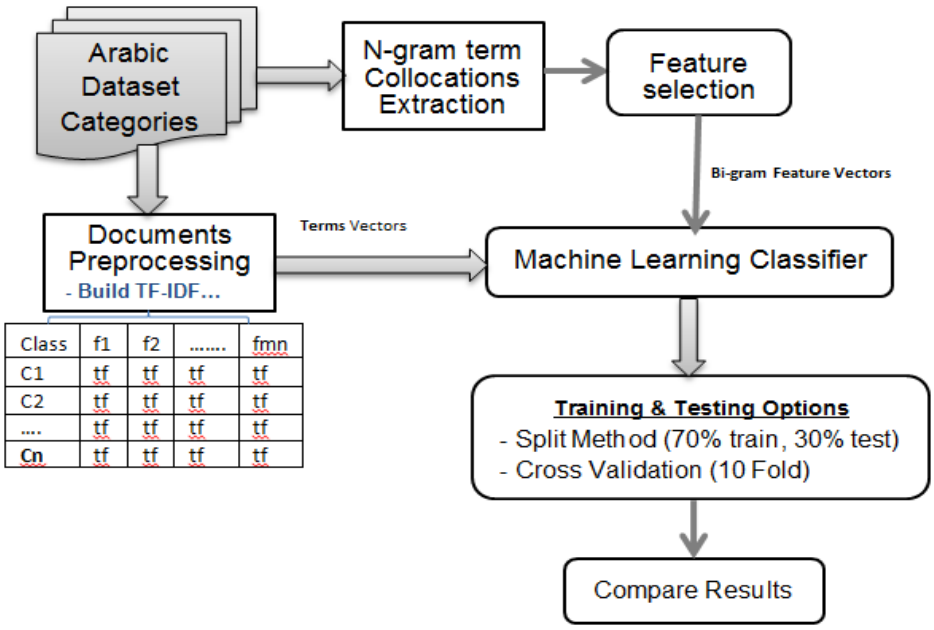

Figure 1: Proposed approach Methodology

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

a. Documents Preprocessing: This step contains three processes:

- Tokenization, for each document in the Arabic corpus: this process is responsible for breaking the Arabic sentences into tokens.
- Stop-words removal, for the tokens generated by the Tokenization process: this process removes the useless words like "من" (from), "على" (on), etc. Some words appear in the sentences and don't have any meaning or indications about the content such as (so لذلك, withبالإشارة , forبالنسبة) or appearing frequently in the document like pronouns such as (he هو, she هي, they هم). Although the prepositions like (from من, to الى, in في, about عن) or demonstratives like (this هذا, these هؤلاء, those اولئك) or interrogatives like (where اين, which اي, who من). These words may have a bad effect on document classification [2].
- Computing the term weights: this part calculates the terms weights based the Term-Frequency (TF) and Inverse-Document-Frequency (IDF) algebraic measures [23]. Equation (1) displays the computation of TF, equation (2) calculates IDF, and finally equation (3) shows how the term weight has computed.

$$TF_{i,j} = \begin{cases} 1 + \log_2 f_{i,j} & if \quad f_{i,j} > 0 \\ 0 & otherwise \end{cases} \quad (1)$$

$$IDF_i = \log_2 \frac{N}{n_i} \quad (2)$$

$$w_{i,j} = \begin{cases} TF_{i,j} \; x \; IDF_i & if \; f_{i,j} > 0 \\ 0 & otherwise \end{cases} \quad (3)$$

Where TF (I,J) denotes the normalized term frequency for term I in document j, and f (I,J) is the number of occurrences for term I, N denotes the number of the documents in the dataset, and n (I) is the total number of occurrences of term I in all documents.

The indexing module indexes the document terms of the corpus, where the index contains the term weights of words using vector space model, which the documents vectors dj = (⟦ ω⟧_1j,⟦ ω⟧_2j,⟦ ω⟧_3j, …, ⟦ ω⟧_nj), where ⟦ ω⟧_1j is the weight for term i in document j. Therefore, in the Arabic term collocations, the bi-gram term collocations are extracted from the Arabic corpus.

b. N-gram term collocation Extraction: For each document in the Arabic dataset, this stage is responsible for extracting the bi-gram term collocations based on the bi-gram method [16], and using enhanced Mutual Information (EMI) measures to score the extracted collocations [1].

c. Feature Selection: this process is responsible for selecting a subset of related features (variables, predictors) to be used in model construction. Many feature selection techniques were used in many researches (such as TF, TF×IDF, CHI square, etc.). The information gain is used in this approach. It is an attribute tilled how much information with respect to the classification target the attribute gives. That is, it measures the difference in information between the cases where you know the value of the attribute and where you do not know the value of the attribute. However, information gain can also be define with mutual information. In particular, A, as shows, in equation (4) information gain IG. (A) Is the reduction in the entropy that is archived by learning a variable?

$$IG(A) = H(S) - \sum_i \frac{S_i}{S} H(S_i)$$

Where H(S) is the entropy of the given dataset and H (Si) is the entropy of the ith subset generated by partitioning S based on feature A.

d. Machine learning Classifier: In this part, we will present an illustrative example of a classifier-based study such as Vector space model (SVM), Naïve Bayesian (NB), and KNN.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1. Extract Term Collocation

This section describes the paper main goal, which refers to the extraction of the term collocations automatically from the dataset/corpus. The bi-gram term collocations are extracting from the Arabic corpus to be exploited in the query expansion process. Equation 5 shows the scoring bi-gram term collocations according to the (EMI) [23].

$$\text{EMI-score} = \log_2 \frac{P(x.y)}{P(x)\,P(y)} * log f(x,y) \quad (5)$$

Where x denotes first term, y denotes second term, P(x) is the occurring probability of first term, P(y) is the occurring probability of second term, P(x,y) is the occurring join probability for first term and second term together, and finally f(x,y) is the frequency of first term and second term together.

In the following, both term collocations extraction and sample examples are described in details. We assume that the algorithm extracts and scores the bi-gram collocations for the term "قطاع", which are the bi-gram collocations shows in the following Table 1:

Table.1 sample collocations for word "قطاع"

Word	Preceding word Freq.	Subsequence word Freq.	EMI
الاعمال	0	8	2.7611
التامين	1	6	2.5971
الخدمات	1	5	2.3086
البنوك	0	6	2.3065
الشحن	0	5	2.2142
الصناعة	0	5	2.2084
احتل	4	0	1.9936
واخيراً	4	0	1.9897
احتياجات	4	0	1.9717
المرتبة	4	0	1.9665
البتروكيماويات	0	4	1.9639
الاستثمار	0	4	1.8287

4. EXPERIMENTS AND EVALUATION RESULTS

To measure the effectiveness and accuracy of both the proposed approach and the full-corpus vs. the summary-corpus method, we have conducted three experiments. In these experiments, the four categories (Economy/ business, Politics, Religion and Science) listed in table 2 have been tested. These experiments are comparing the proposed approach method vs. the full-corpus, comparing the proposed approach versus the summary-corpus method. The approach, as presented in Figure 1, is to explain classifier algorithm where document contents are associated to a category, by extracting the vital terms that represent the syntax of a document by virtue of feature represented by vector space techniques.

A. Dataset Collection

The dataset collection/corpus that we used consists of 800 Arabic text documents divided into four categories Economy, Politics, Religion, and Science, 200 documents for each class as show in Table 2. All documents are classified of short articles. It is a subset of 60913-document corpus collected from many newspapers and other web sites. The text documents have been preprocessed before being used, each document have has been tokenized, i.e. split into tokens according to the white space position.

The experiments conducted using Waikato Environment for Knowledge Acquisition (WEKA) [22], where SVM, J48, and NB are already implemented.

Table 2: Data Collection Statistics

Category	#of total Documents	# of total words in class	# of total unique words in class
Economic (اقتصاد)	200	75,334	21,291
Politic (سياسة)	200	78,636	27,465
Religion) (دين	200	52,804	17,794
Science (علوم)	200	58,068	13,950
Total	**800**	**264,842**	**62,177**

Figure 2 shows the snapshot of sample for frequency (W_F) and document frequency (D_F) for each word in the corpus in class.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Word	W_F	D_F	Classes_Name
العالميه	239	94	Economy , Politics ,
رابين	238	100	Economy , Politics ,
بني	232	86	Economy , Politics , Religion
واحد	214	141	Economy , Politics , Religion
العمل	212	103	Economy , Politics , Religion
فاذا	212	157	Economy , Politics , Religion
مليون	204	70	Economy , Politics ,
يعرض	201	107	Economy , Politics , Religion
الوجع	196	83	Politics , Science
العربي	194	79	Economy , Politics , Science
الصداع	193	32	Science
الاقتصاديه	190	78	Economy , Politics ,
المسيح	188	53	Politics , Religion ,
الملك	187	88	Economy , Politics , Religion
البدن	187	100	Religion , Science
الدماغ	187	77	Science
اجل	186	122	Economy , Politics , Religion
عمر	184	86	Economy , Politics , Religion
عدد	183	102	Economy , Politics , Religion
ما	182	112	Economy , Politics , Religion
يقول	182	132	Economy , Politics , Religion

Figure 2: snapshot of the sample frequency and classes' name

B. Tools Description

The methodology consists of selecting available open source data-mining tools to test our approach. Many open data mining tools are available on the Web and free of cost. The tool chosen is the Waikato Environment for Knowledge Analysis (WEKA). The WEKA toolkit [22] is a widely used toolkit for machine learning and data mining that were originally developed at the University of Waikato in New Zealand.

C. Evaluation Mertic

The dataset are test using percentage split method, where 70% of the data used as a training and the remaining 30% are use as testing. The performance of the classifier (in classifying the full and the summarized documents) is measured with respect to the accuracy, precision, and recall. Additionally, the time to construct the model is included in the evaluation analysis, which can be measure by the following equations:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

Where TP true positive rate, FP the false positive rate, TN the true negative rate, and FN the false negative rate.

D. Text Sumerization

Many feature selection techniques were used in many researches (such as TF, TF×IDF, CHI square, etc.). Then the documents summaries can be used as inputs to machine learning systems rather than full-text documents, we proposed to use text summarization as a feature selection for classifying Arabic documents using SVM [21].
In addition, Text summarization is the process in which a computer receives a text document(s) as input and produces a summary of that document(s) as an output. It makes it easy to extract just the important sentences within a document by highlighting the best-related sentences within a text. Key words extractor and spelling corrector are used in forming the summary. In this research, we used the texts that are used by the Sakhr summarizer [20] to summarizing the Arabic documents. To test the accuracy of the proposed approach, it is comparing to the classification results of these texts [21].

E. Expemental Results

In these experiments, we have used three of different of dataset/corpus, which are full-corpus, Summery-corpus, and full-corpus-bigram (our-approach). The full-corpus is the dataset/documents without preprocessing. The summery-corpus is the dataset/documents which are summarized by using sakhr summery. And, the full-corpus-bigram is the dataset/documents which are preprocessing by using bi-gram for every two terms assigned together. To test the accuracy of the proposed approach, it's comparing to the performance of the classifiers for these corpus. Compare our approach (full-corpus-bigram) with the summery approach and the original corpus (full-corpus). Additionally, the four-classifier algorithm SVM, 48J, KNN, and NB have compared with the dataset used.

Table 3 shows the comparison of the experimental results by the proposed approach, the full-corpus, and the summery-corpus in terms of precision and Time taken to test model.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 3: The comparison of the experimental results

		Dataset/Corpus		
Algorithm	**Measure**	Full-Corpus	Summery-Corpus	**Our approach** (Full-Corpus-Bigram)
SVM	Precision	**0.98**	0.932	**0.993**
Naïve Bayesian	Precision	0.98	0.95	0.983
J48	Precision	0.93	0.81	0.99
KNN	Precision	0.86	0.47	0.99

Table 4 shows the comparison of the experimental results of the executed time needs, and Table 5 shows the comparison of the algorithm accuracy by the proposed approach, the full-corpus, and the summery-corpus.

Table 4: The comparison of the experimental results of Execution Time needs

	Dataset/Corpus		
Algorithm	Full-Corpus	Summery-Corpus	**Our approach** (Full-Corpus-Bigram)
SVM	0.35 seconds	0.44 seconds	0.14 seconds
Naïve Bayesian	2.01 seconds	1.73 seconds	0.74 seconds
J48	4.98 seconds	10.81 seconds	0.55 seconds
KNN	0.75 seconds	0.39 seconds	0.36 seconds

Table 5: The comparison of the accuracy for the experimental results

	Algorithm/ Accuracy %			
Dataset/Corpus	**SVM**	**Naïve Bayesian**	**J48**	**KNN**
Full-Corpus	93	98	93	81.6
Summery-Corpus	94	95	78.9	43
Our approach (full-Corpus-Bigram)	99.5	99.3	99.1	99.5

Figure 3 shows the precision values of the proposed approach versus the full-corpus based method and summery-corpus. For all the tested algorithms, the precision values of the proposed approach are higher than or nearly to the precision values of the full-corpus and summery-corpus method.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

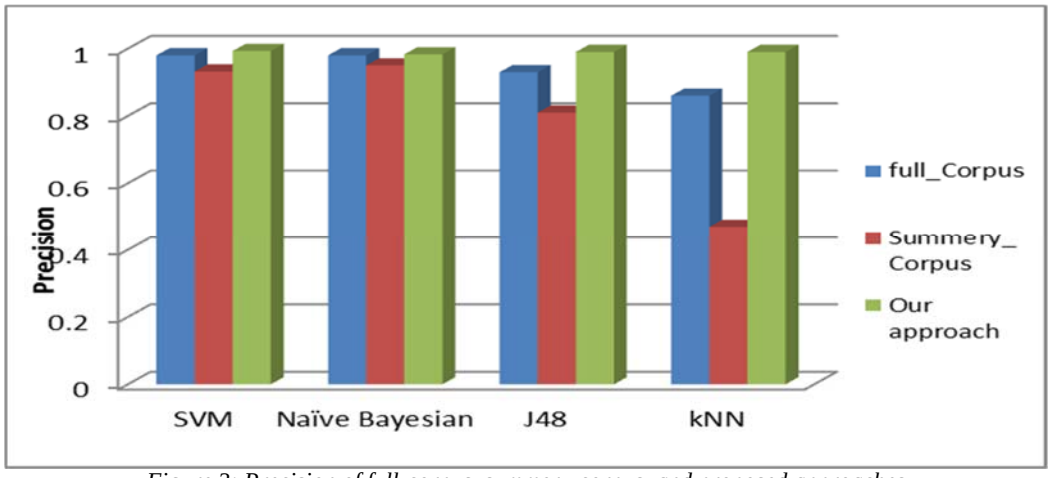

Figure 3: Precision of full-corpus, summery-corpus, and proposed approaches

Figure4 shows the time taken to test the model in the tool used of the proposed approach versus the full-corpus based method and summery-corpus. For all the tested algorithms, the time taken values of the proposed approach are less than to the time in the second values of the full-corpus and summery-corpus method.

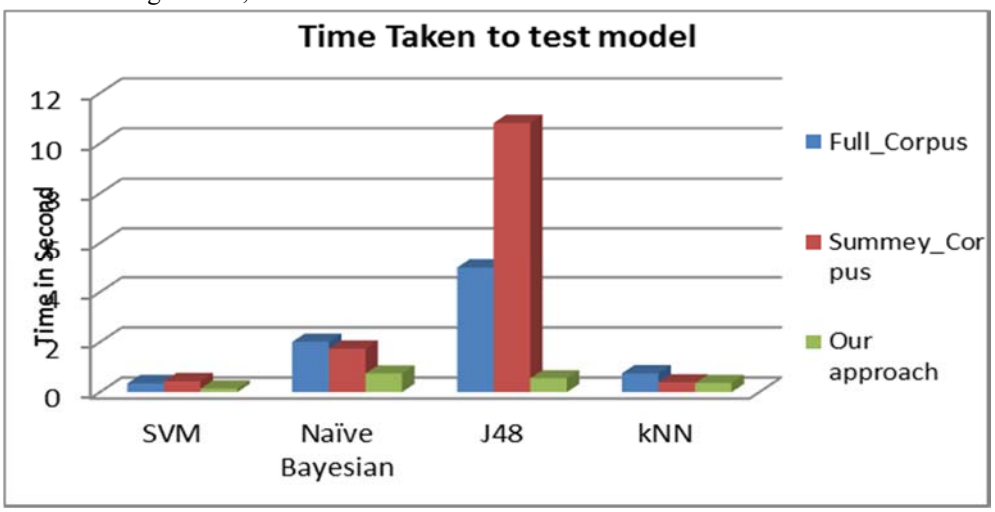

Figure 4: Accuracy for full-corpus, summery-corpus and proposed approaches

Table 6: A comparison between our proposed and the related work

#	Research	Category	Technique	Dataset/Corpus	Algorithms	Avg.Accuracy%
1	Elthweb [21]	Classification	Summarization	Arabic Docs. 4 classes	SVM, NB	92
2	Alshalabi et. Al [8]	Classification	N-gram and single terms	Arabic Docs.	KNN	87
3	In paper [17]	Classification	Single terms	Arabic Docs	KNN	82
4	**Our approach**	Classification	**Terms Collocations**	Arabic Docs. 4 classes	SVM, NB,J48,KNN	**99.5**

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. CONCLUSION

In this paper, we empirically studied and tested the categorization effectiveness of using automatic machine learning Techniques (AMLT) on Arabic text categorization based on term collocation. Selection techniques (SVM, J48, NB and KNN) are applied on the Arabic corpus using bi-gram term collocations. The experiments on an Arabic corpus have demonstrated that the automatically extracted text from the Arabic corpus is more effective in representing and classifying Arabic documents. Therefore, there is no need for any preprocessing for the text before such as stemming, summarization process. The results indicated that (AMLT) using features term collection has better accuracy than the full-corpus and summery-corpus methods. In addition, the results indicated that our approach outperform the full-corpus and summery-corpus method in all used measures. The result for evaluation shows that the (AMLT) is capable of classifying word semantics, which can facilitate semantic analysis of Arabic text classification.

ACKNOWLEDGE:

This research has been funded by the deanship of scientific research, Najran University-Saudi Arabia [research Grant: NU/ESCI/15/076].

REFERENCES:

[1] A. M. Saif and M. J. A. Aziz, "An Automatic Collocation Extraction from Arabic Corpus," Journal of Computer Science, vol. 7, no. 1, pp. 6 - 11, 2011.

[2] B., Al-Shargabi, F., Olayah, and W., Alromimah,"An Experimental Study for the Effect of Stop Words Elimination for Arabic Text Classification Algorithms" International Journal of Information Technology and Web Engineering (IJITWE) 6(2), 68-75, April-June 2011.

[3] S. Al-Saleem, "Automated Arabic Text Categorization Using SVM and NB", International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011.

[4] Beal V. Search Engine. 2011. Available at: www.webopedia.com/TERM/S/search_engine .html (accessed April 2015).

[5] Shaalan K and Oudah M. A hybrid approach to Arabic named entity recognition. Journal of Information Science (JIS) 2014; 40: 67-87.

[6] Group M. Middle East Internet Usage & Population Statistics. 2015. Available at: www.internetworldstats.com/stats5.htm (accessed May 2017).

[7] Al-Zoghby A, Sharaf A and Hamza T. Arabic Semantic Web Applications–A Survey. Journal of Emerging Technologies in Web Intelligence (JETWI) 2013; 5: 52-69

[8] Al-Shalabi, R. and Obeidat, R., 2008, March. Improving KNN Arabic text classification with n-grams based document indexing. In Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt (pp. 108-112).

[9] Darwish, K. and Magdy, W., 2014. Arabic information retrieval. Foundations and Trends® in Information Retrieval, 7(4), pp.239-342.

[10] A., Farghaly, and Shaalan K., "Arabic Natural Language Processing: Challenges and Solutions", ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, no. 4, New York, NY, USA, ACM, pp. 1-22, 2009.

[11] R., Al-Shalabi, Kanaan G, " Proper Noun Extraction Algorithm for Arabic Language" international Conference on IT to celebrate S. Charmonment's 72nd Birthad. March 2009, Thailand

[12] Alromima, Waseem, Ibrahim F. Moawad, Rania Elgohary, and Mostafa Aref. "Extracting N-gram terms collocation from tagged Arabic corpus." 9th International Conference on Informatics and Systems (INFOS), pp. NLP-10. IEEE, 2014.

[13] Al-Shalabi, R.,Kanaan, G., Jaam, J.M.. Hasnah, A. and Hilat, E. 2004. Stop-word Removal Algorithm for Arabic Language. Proceedings of 1st International Conference on Information & Communication Technologies: from Theory to Applications, CTTA'04, (Damascus, Syria, April 2004). IEEE-France, 545-

[14] K., AlKhateeb and A. Badarneh, "Automatic extraction of Arabic multi-word terms," in Computer Science and Information Technology (IMCSIT), Proceedings of the 2010 International Multiconference on, Wisla, 2010.

[15] Gharib, T.F, and Badieh H.M, Arabic Text Classification Using Support Vector Machines, International Journal of Computers and Their Applications, 16, 4, 2009

[16] B. S., B. Daille and D. Aboutajdine, "A multi-word term extraction program for Arabic language," in Proceeding of the 6th

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

International Conference on Language Resources and Evaluation, Morocco, 2008.

[17] A. Mesleh, "Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System", Journal of Computer Science (3:6), pp. 430-435, 2007.

[18] Koulali, R., El-Haj, M. and Meziane, A., 2013, May. Arabic Topic Detection using automatic text summarisation. In Computer Systems and Applications (AICCSA), 2013 ACS International Conference on (pp. 1-4). IEEE.

[19] Abainia, K., Ouamour, S. and Sayoud, H., 2015, November. Neural Text Categorizer for topic identification of noisy Arabic Texts. In Computer Systems and Applications (AICCSA), 2015 IEEE/ACS 12th International Conference of (pp. 1-8). IEEE.

[20] Sakhr company website: http//:www.sakhr.com, text summerization.

[21] Al-Thwaib, E., 2014. Text summarization as feature selection for arabic text classification. World of Computer Science and Information Technology Journal (WCSIT), 4(7), pp.101-104.

[22] WEKA. Data Mining Software in Java: http://www.cs.waikato.ac.nz/ml/weka. Last visit on May, 2017.

[23] Baeza-Yates R and Ribeiro-Neto B. Modern information retrieval the Concepts and Technology behind Search. 2ed edn.:Addison-Wesley, 1998.

[24] Ali, Imran., 2012. Application of a mining algorithm to finding frequent patterns in a text corpus: A case study of the arabic. International Journal of Software Engineering and Its Applications, 6(3), pp.127-134.

[25] Zhang, C., Wang, H., Cao, L., Wang. W. and Xu, F., 2016. A hybrid term–term relations analysis approach for topic detection.. Knowledge-Based Systems, Issue 93, pp. pp.109-120

[26] Group M. Middle East Internet Usage & Population Statistics. 2015. Available at: www.internetworldstats.com/stats5.htm (accessed May 2017).

[27] Al Khatib, K., & Badarneh, A. (2010). Automatic extraction of Arabic multi-word terms. In proceedings: of the 2010 International Multiconference on Computer Science and Information Technology (IMCSIT), pp. 411–418.