Big Data Categorization for Arabic Text Using Latent Semantic Indexing and Clustering

Fawaz S. Al-Anzi[1] and Dia AbuZeina[2]

***Abstract*—**Documents categorization is an important field in the area of natural language processing. In this paper, we propose using Latent Semantic Indexing (LSI), singular value decomposing (SVD) method, and clustering techniques to group similar unlabeled document into pre-specified number of topics. The generated groups are then categorized using a suitable label. For clustering, we used Expectation–Maximization (EM), Self-Organizing Map (SOM), and K-Means algorithms. In our experiments, we use a corpus that contains **1**000 documents of ten topics (100 document for each topic. The results show that using LSI and clustering techniques for Arabic text categorization achieves good performance. The results show that EM clustering method outperforms other investigated clustering methods with an average categorization accuracy of 89%.

***Keywords*—**Arabic Text, Categorization, Classification, Clustering, Unsupervised Learning.

I. Introduction

THE significant growth of online textual information has increased the demand for effective content-based text retrieval methods. Document clustering approaches have grown significantly to fulfil a wide range of applications in different fields. Such applications include an efficient and effective methods for matching the interests in the Social networks, such as Facebook, Twitter and Google+ as in [1]. In this work, we utilized document-clustering techniques for Arabic text categorization. Text categorization is the activity of labelling natural language texts with thematic categories from a predefined set as in [2].

In this paper, we proposed a novel approach to categorize Arabic text using unsupervised Machine Learning (ML) methods. The proposed method utilize the Latent Semantic Indexing (LSI) technique to generate the required feature using singular valued decomposition (SVD) method. The features are then clustered into a predefined number of groups representing the similar documents. Clustering methods include Expectation–Maximization (EM), Self-Organizing Map (SOM), and K-Means algorithms. There are many ML tools such as WEKA, Rapid Minor, and ORANGE. In this work, we use WEKA.

In the next section, we present literate review. In section III, we present the literature for text categorization in English and other languages. We present the Arabic language challenges in section IV. Then, the proposed method are presented in section V, followed by results in section VI. The conclusion and future work are presented in section VII.

II. Literature Review

In this section, we present a summary of the research literature on Arabic text categorization. There are quite some research work based on supervised approach for flat corpuses. However, the authors could not locate any research work regarding hierarchical (multi-level categorization). This literature review focuses on Arabic flat documents categorization only as we have not found any research work related to hierarchical Arabic text categorization. For supervised approaches, a comparison was performed to compare the performance of three classifiers Support Vector Machines (SVM), Naïve Bayes (NB) and Decision Trees (DT) classifiers as in [3]. For unsupervised approaches, a PhD dissertation was presented for some Information Retrieval (IR) related topics such as light stemming, weighting scheme, LSI, cosine diffusion map space, and SVD as in [4]. Reference [5] showed that clustering documents have significant results compared with traditional information retrieval system without clustering.

III. Document Clustering for English and Other Languages

Unsupervised approaches for text classification and documents clustering have been widely investigated in English and other languages. Reference [6] introduced an approach which uses frequent item (term) sets for text clustering. Reference [7] describe two flat clustering algorithms: the K-Means algorithm, an efficient and widely used document clustering method, and the expectation-maximization algorithm, which is computationally more expensive, but also more flexible. Reference [8] presented an approach for clustering scientific documents based on the utilization of citation contexts. Reference [9] gave a brief overview of the document clustering research and the developments in this field. Reference [10] discussed the issues that need to be addressed in the development of a web-clustering engine including acquisition and preprocessing of search results and their clustering and visualization.

Fawaz S. Al-Anzi[1] is with department of computer engineering, Kuwait University, 13060, Kuwait.

Dia AbuZeina[2], is with Research Sector, Kuwait University13060, Kuwait.

http://dx.doi.org/10.15242/IIE.E0116001

IV. Arabic Text Categorization Challenges

In the recent years, there has been an increasing interest in Arabic Natural Language Processing (NLP) research. Arabic is the native language of population of more than 380 million and extends over large geographical areas in northern Africa and Middle East as in [11]. From NLP point of view, Arabic is characterized by a number of challenges that need to be considered when developing Arabic NLP applications. For example, the direction of writing, diacritized and non-diacritized text, not using of capitalization for proper nouns, etc. Many researches addressed Arabic NLP challenges. Reference [12] listed six of Arabic difficulty sources.

As it is known, modern written Arabic language is usually undiacritized and the reader can seamlessly read undiacritized text. However, this character might present challenges in text categorization. For illustration, we present some examples of short sentence obtained from the Holy Quran. Even though the Holy Quran is diacritized, we present these sentences for illustration purpose. Table I shows two short sentences that belong to completely different categories. The listed sentences were intentionally chosen to have exactly same word with different meaning. Intuitively, any text categorization algorithm is expected to correctly assign the class for each sentence regardless of the ambiguity that might be raised due to the context. Robust text categorization algorithm should address this ambiguity by assigning the meaning word "paradise" category to the first sentence and the meaning word "lunacy" to the second one. The material used in Table I was obtained from the website [13]. The items in Table I is organized as the following: the source of the Quranic sentence, the script in Arabic, the pronunciation of the sentence, and the meaning. The number besides the pronunciation and meaning is the sentence number where it was found in the Sūra. In Table I, the word we chose as an example is "جنة" which has different pronunciations according to the diacritization.

TABLE I
Ambiguity In Arabic Text Categorization

#	Example	category
1	Sūra XXVI.: Shu'arāa, or The Poets وَٱجۡعَلۡنِي مِن وَرَثَةِ جَنَّةِ ٱلنَّعِيمِ 85. WaijAAalnee min warathati jannati alnnaAAeemi 85. Make me one of the inheritors Of the Garden of Bliss	الجنة Paradise (Garden)
2	Sūra VII.: A'rāf, or the Heights أَوَلَمۡ يَتَفَكَّرُواْۗ مَا بِصَاحِبِهِم مِّن جِنَّةٍۚ إِنۡ هُوَ إِلَّا نَذِيرٞ مُّبِينٌ 184. Awa lam yatafakkaroo ma bisahibihim min jinnatin in huwa illa natheerun mubeenun 184. Do they not reflect? Their Companion is not seized With madness: he is but A perspicuous warner.	جنون Lunacy (Madness)

We also present another challenge regarding Arabic text categorization which is stemming. Even though many researches used stemming as a dimension reduction method, we present some sentences from the Holy Quran that show stemming might not always be good option for text categorization. Table II shows a number of different words that has same root that belong to different categories. Hence, stemming will use same root for completely different words, which may lead to performance reduction. The root was obtained from the website of Reference [14] which also has indicated the root (جنن) has 46 distinct words of a total 201 words. The information in the Table II is arranged as the name of the Quranic Sūra where the sentence has been found, the written Arabic text of the sentence, the pronunciation, and the meaning.

TABLE II
Same Roots For Different Words In Arabic Text

#	Example	root
1	Sūra LVI.: Wāqi'a, or The Inevitable Event فَرَوۡحٞ وَرَيۡحَانٞ وَجَنَّتُ نَعِيمٖ 89. Farawhun warayhanun wajannatu naAAeemin 89. (There is for him) Rest And atisfaction, and A Garden of Delights.	جنن "Paradise Graden"
2	Sūra LIII.: Najm, or the Star هُوَ أَعۡلَمُ بِكُمۡ إِذۡ أَنشَأَكُم مِّنَ ٱلۡأَرۡضِ وَإِذۡ أَنتُمۡ أَجِنَّةٞ فِي بُطُونِ أُمَّهَٰتِكُمۡۖ فَلَا تُزَكُّوٓاْ أَنفُسَكُمۡۖ هُوَ أَعۡلَمُ بِمَنِ ٱتَّقَىٰٓ 32. huwa aAAlamu bikum ith anshaakum mina al-ardi wa-ith antum ajinnatun fee butooni ommahatikum fala tuzakkoo anfusakum huwa aAAlamu bimani ittaqa 32. He knows You well when He brings You out of the earth, And when ye are hidden In your mothers' wombs. Therefore justify not yourselves; He knows best who it is That guards against evil.	جنن "Foetuses"

In the literature, we found a debate among researchers about the benefits of using stemming in Arabic text categorization. Reference [12] indicated that no stemming is applied, because it is not always beneficial for Arabic text categorization tasks, since many terms may be conflated to the same root form. Reference [15] used stemming, but there was no baseline to investigate the effect of stemming.

V. The Proposed Method

LSI is used to generate SVD by means of a linear algebra matrix decomposition technique. SVD is based on a theorem from linear algebra which says that a rectangular m-by-n matrix A can be broken down into the product of three matrices - an orthogonal matrix U, a diagonal matrix S, and the transpose of an orthogonal matrix V . The theorem is usually presented something like this: *Amn = UmmSmnVTnn*. Fig. 1 demonstrates the reduced rank SVD. The black bold line in the shaded region of S represents the values retained in computing rank k approximation. In general, we do not take all singular values; instead we only consider the most important values starting from the first singular values up to the desired value. There are many free and commercial software available for related LSI. We used MATLAB for decomposing A. Genism is a Python-based package that can be used to compute SVD as in [16].

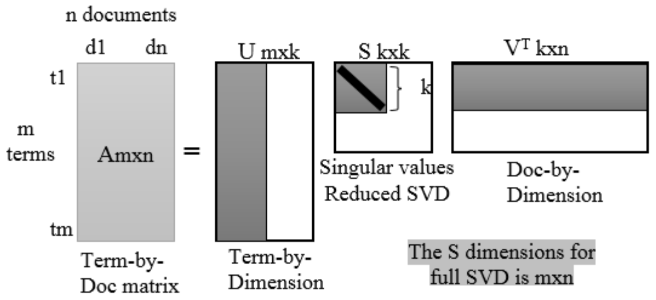

Fig. 1 SVD representation of the document-term(s) matrix

In Fig. 1, the Doc-by-Dimension matrix is also called document-vectors which we we focused on in our proposed method. The proposed method is the process of clustering the document-vectors. The purpose of clustering is to partition a give data set into a pre specified number of groups such that data in a particular cluster are more similar to each other than object in different clusters as in [17].

Regarding the corpus used in our experiments, we created a corpus that contains 1000 documents belonging to 10 different categories. The corpus contains more than 500,000 words that include more than 100,000 unique words. We got the documents from Alanba newspaper website in Kuwait [18]. It provides an access for a large number of documents which are organized based on the categories. The corpus topics include 100 document of each of the following categories: Health, Economy, Sports, Islam and Sharia, Education, Arts and Artists, Municipality affairs, Tourism and travel, Security and law, and Technology.

The proposed method is summarized using the following algorithm:

Step 1: From the corpus, a term_by_document matrix is created using only term counts. Before running the code, we set the following:

The stop list and the ignore characters are specified.

The document frequency is set (we propose DF to be set >=13).

The short words are also declared (we propose to exclude words less than 4 characters).

Finally, the term_by_document matrix is weighted using TF*IDF.

Step 2: The created term_by_document matrix is then used to compute SVD. The rank k approximation is set. We propose to investigate different k values around k=30.

Step 3: The document-vector matrix created in step 2 is clustered using a clustering algorithm. The number of clusters is set to 10 since we have a corpus of 10 categories.

Step 4: For evaluation, an accuracy measure can be used which is the rate of correctly predicted topics. The accuracy is defined as : *Accuracy* = (TP + TN)/N, where N = TP + FP + TN + FN, where TP is True positive, FP False positive, TN True negative, FN False negative as in [19].

VI. The Experimental Results

In this section, we present the experimental results. We follow the steps described in section V, the proposed method. The stop list is specified as : { , الكويتيين , الكويتي , الكويتية , الكويت الأنباء , الانباء , بالكويت }. The ignore characters list is: {~, ` , ! , @ , # , £ , € , $, % , ° , ^ , & , * , (,) , - , _ , + , = , » , « , { , } , [,] , | , \ , / , : , ; , 0,1,2,3,4,5,6,7,8,9}. The document frequency threshold is assigned 13 and all words that are less than 4 characters in lengths are ignored.

We performed our experiments in 2 cases; 5 categories and 10 categories. The results in Table III show that the accuracy of 5 categories is the best in two cases when k=18 and when k=30. The EM algorithm was applied. The 5 categories included are {Health (HE), Economy (EC), Sports (SP), Islam and Sharia (IS), Education (ED)}.

TABLE III
The Clustering Performance of Different k Values

Singular Value	Category Accuracies	Avg Accuracy

k=6	HE=70%, EC=97%,
	SP=100%, IS=98%,
	ED=98%
		92.6%

k=12	HE=70%, EC=94%
	SP=100%, IS=98%
	ED=100%
		92.4%

k=18	HE=100% , EC=92%
	SP=100% , IS=96%
	ED=97%
		97%

k=24	HE=100% , EC=93%
	SP=100% , IS=97%
	ED=93%
		96.9%

k=30	HE=100% , EC=94%,
	SP=100% , IS=97%,
	ED=94%
		97%

k=36	HE=63% , EC=95%
	SP=100% , IS=98%
	ED=95%
		90.2%

k=42	HE=97% , EC=99%
	SP=88% , IS=96%
	ED=93%
		88.6%

k=48	HE=90% , EC=94%
	SP=100% , IS=72%
	ED=92%
		89.6%

In Table III, we found that the high accuracy was obtained in the range of singular values starting at k=18, 24, and 30.

For the SOM and K-Means, we used k=24 as it was in the middle of k values. The categorization accuracy using SOM was 93.4% and it was extremely low using K-Means. We also performed an experiment for the entire corpus (ten categories) for different values of k, k=18, 24, 30, and 36. We have got the maximum accuracy 89.1% when k=30.

For labeling, we chose among the high frequency 50 words that are more than four characters in length. In fact, we don't need the small words to dominate the cluster name. For Islamic and sharia we got the list :{ , العمل , رمضان , الزكاة , تعالى النبي , القرآن , الناس }. According to google translator, the previous list is translated to: {almighty, Zakat, Ramadan, work, people, the Koran, the Prophet} [20]. No doubt that the generated words are belong to Islamic and sharia documents. Hence, we can generate the cluster label that includes all these keywords that concatenate by appropriate connective (-, &, etc), or use a single word such as “دين”, religion. The same method can be used for other clusters.

VII. Conclusion and Future Work

In this work, we proposed Arabic text categorization method for Big Data environment. We employed LSI and some of clustering techniques for this task. The results showed that this technique is an excellent approach to label documents without training data. As future work, we hope to be able to research and investigate the clustering accuracy without previously specifying the number of clusters.

VIII. Acknowledgment

This work is supported by Kuwait Foundation of Advancement of Science (KFAS), Research Grant Number P11418EO01 and Kuwait University Research Administration Research Project Number EO06/12.

We would like to acknowledge the convenience that this research got from using Alanba newspaper website in Kuwait as a source of Arabic text documents.

References

[1] Poomagal, Shanmugam, Palanisamy Visalakshi, and Thiagarajan Hamsapriya. "A novel method for clustering tweets in Twitter." International Journal of Web Based Communities 11.2 (2015): 170-187.
http://dx.doi.org/10.1504/IJWBC.2015.068540

[2] Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1 (2002): 1-47.

[3] Khorsheed, Mohammad S., and Abdulmohsen O. Al-Thubaity. "Comparative evaluation of text classification techniques using a large diverse Arabic dataset." Language resources and evaluation 47.2 (2013): 513-538.
http://dx.doi.org/10.1007/s10579-013-9221-8

[4] Ataa Allah, Fadoua. "Information retrieval: applications to English and Arabic documents." (2008).

[5] Ghwanmeh, Sameh H. "Applying Clustering of hierarchical K-means-like Algorithm on Arabic Language." International Journal of Information Technology3.3 (2005).

[6] Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text clustering." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.
http://dx.doi.org/10.1145/775047.775110

[7] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze.Introduction to information retrieval. Vol. 1. Cambridge: Cambridge university press, 2008.
http://dx.doi.org/10.1017/CBO9780511809071

[8] Aljaber, B., Stokes, N., Bailey, J., & Pei, J. (2010). Document clustering of scientific texts using citation contexts. Information Retrieval, 13(2), 101-131.
http://dx.doi.org/10.1007/s10791-009-9108-x

[9] Andrews, Nicholas O., and Edward A. Fox. "Recent developments in document clustering." (2007).

[10] Carpineto, Claudio, et al. "A survey of web clustering engines." ACM Computing Surveys (CSUR) 41.3 (2009): 17.

[11] Mubarak, Hamdy, and Kareem Darwish. "Using Twitter to collect a multi-dialectal corpus of Arabic." ANLP 2014 (2014): 1.
http://dx.doi.org/10.3115/v1/w14-3601

[12] Moh’d Mesleh, Abdelwadood. "Feature sub-set selection metrics for Arabic text classification." Pattern Recognition Letters 32.14 (2011): 1922-1929.
http://dx.doi.org/10.1016/j.patrec.2011.07.010

[13] The Holy Qur'an. (2015, August). Retrieved from http://www.sacred-texts.com/isl/quran/

[14] Almaany. (2015, August). Retrieved from http://www.almaany.com/quran-b/

[15] Harrag, Fouzi, Eyas El-Qawasmah, and Abdul Malik S. Al-Salman. "Comparing dimension reduction techniques for Arabic text classification using BPNN algorithm." Integrated Intelligent Computing (ICIIC), 2010 First International Conference on. IEEE, 2010.
http://dx.doi.org/10.1109/iciic.2010.23

[16] Gensim. (2015, August). Retrieved from https://radimrehurek.com/gensim/index.html
Jain, Anil K., and Richard C. Dubes. Algorithms for clustering data. Prentice-Hall, Inc., 1988.

[17] Alanba. (2015, August). Retrieved from http://www.alanba.com.kw/newspaper/

[18] Yiming Yang and Thorsten Joachims (2008) Text categorization. Scholarpedia, 3(5):4242.
http://dx.doi.org/10.4249/scholarpedia.4242

[19] Google Translate. (2015, August). Retrieved from https://translate.google.com/