Vector Space Models to Classify Arabic Text

[1]Jafar Ababneh, [1]Omar Almomani,, [2]Wael Hadi, [1]Nidhal Kamel Taha El-Omari, and Ali Al-Ibrahim

[1]Department of Network and Computer Information System, Faculty of Information and Technology/ The World Islamic Sciences & Education University, Amman, Jordan

[2]Department of Management Information System, Faculty of Administrative & Financial Sciences, Petra University, Amman, Jordan

ABSTRACT-- *Text classification is one of the most important tasks in data mining. This paper investigates different variations of vector space models (VSMs) using KNN algorithm. The bases of our comparison are the most popular text evaluation measures. The Experimental results against the Saudi data sets reveal that Cosine outperformed Dice and Jaccard coefficients*

Keywords: *Arabic data sets, Data mining, Text categorization, Term weighting, VSM.*

I. INTRODUCTION

Text categorization (TC) is one of the most important tasks in information retrieval (IR) and data mining [1]. This is because of the significance of natural language text, the huge amount of text stored on the internet, and the available information libraries and document corpus. Further, TC importance rises up since it concerns with natural language text processing and classification using different techniques, in which it makes the retrieval and other text manipulation processes easy to execute.

Many TC approaches from data mining and machine learning exist such as: decision trees [2], Support Vector Machine (SVM) [3], rule induction [4], Statistical Classification Methods [5], and Neural Network [6]. The goal of this paper is to present and compare results obtained against Arabic text collections using K-Nearest Neighbor algorithm. Particularly, three different experimental runs of the KNN algorithm on the Arabic data sets we consider are performed, using three different VSMs (Cosine, Dice, Jaccard).

Generally, TC based on text similarity goes through two steps: Similarity measurement and classification assignment. Term weighting is one of the known concepts in TC, which can be defined as a factor given to a term in order to reflect the importance of that term. There are many term weighting approaches, including, inverse document frequency (IDF), weighted inverse document frequency (WIDF) and inverse term frequency (ITF) [7]. In this paper, we compare different variations of VSMs with KNN [8] algorithm using IDF. The bases of our comparison between the different implementations of the KNN are the F1, Recall, and Precision measures [9]. In other words, we want to determine the best VSM, which if which if merged with KNN produces good F1, Precision and recall results. To the best of the author's knowledge, there are no comparisons which have been conducted against The Saudi Newspapers (SNP) using VSM.

The organization of this paper is as follows, related works are discussed in Section 2. TC problem is described in Section 3. In Section 4, experiment results are explained, and finally conclusions and future works are given in Section 5.

II. RELATED WORK

There are over 320 million Arabic native speakers in 22 countries located in Asia and Africa [10]. Due to the enormous energy resources, the Arab world has been developing rapidly in almost every sector especially in economics. As a result, a massive number of Arabic text documents have been increasingly arising in public and private sectors, where such documents contain useful information that can be utilized in a decision making process. Therefore, there is a need to investigate new intelligent methods in order to discover useful hidden information from these Arabic text collections.

Reviewing the existing related works proved that there are several methods which have been proposed by researchers towards Arabic text classification. For classifying Arabic text sources the N-Gram Frequency Statistics technique is investigated by [11]. This method is based on both Dice similarity and Manhattan distance measures in classifying an Arabic corpus. For this research the Arabic corpus was obtained from various online Arabic newspapers. The data is associated with four categories. After performing several pre-processing on the data, and experimentation, the results indicated that the "Sport" category outperformed the other categories with respect to recall evaluation measure. The least category was "Economy" with around 40% recall. In general the N-gram Dice similarity measure figures outperformed that of Manhattan distance similarity.

Different variations of Vector Space Model using KNN algorithm were investigated [12], these variations are Cosine coefficient, Dice coefficient and Jacaard coefficient, using different term weighting approaches. The average F1 results obtained against six Arabic data sets indicated that Dice based TF.IDF and Jaccard based TF.IDF outperformed Cosine based TF.IDF, Cosine based WIDF, Cosine based ITF, Cosine based log(1+tf), Dice based WIDF, Dice based ITF, Dice based log(1+tf), Jaccard based WIDF, Jaccard based ITF, and Jaccard based log(1+tf).

III. TEXT CATEGORISATION PROBLEM

TC is the task in which texts are classified into one of predefined categories based on their contents. If the texts are newspaper articles, categories could be, for example, economics, politics, sports, and so on. This task has various applications such as automatic email classification and web-page categorization. Those applications are becoming increasingly important in today's information-oriented society.

TC problem can be defined according to [1] as follows: The documents divided in two datasets, for training and testing. Let training data set = $\{d_1,d_2,\ldots,d_g\}$, where g documents are used as examples for the classifier, and must contain sufficient number of positive examples for all the categories involved. The testing data set $\{d_{g+1},d_{g+2},\ldots,d_n\}$ used to test the classifier effectiveness. The matrix shown in Table 1 represents data splitting into training and testing. A document d_y is considered a positive example to C_k if C_{ky} =1 and a negative example if C_{ky} =0.

Generally, TC task goes through three mainly steps: Data pre-processing, text classification and evaluation. Data preprocessing phase is to make the text documents suitable to train the classifier. Then, the text classifier is constructed and tuned using a text learning approach against from the training data set.

Table 1: Representation of text categorization problem

Category	Training data set			Testing data set		
	d_1	…	d_j	d_{j+1}	…	d_n
C_1	C_{11}	…	C_{1j}	$C_{1(j+1)}$		C_{1n}
…	…	…	…	…	…	…
C_m	C_{m1}	…	C_{mg}	$C_{m(j+1)}$	…	C_{mn}

Finally, the text classifier gets evaluated by some evaluation measures i.e recall, precisinon, etc [9]. The next two sub-sections are devoted to discuss the main phases of the TC problem related to the data we utilised in this paper.

A. DATA PRE-PROCESSING ON ARABIC DATA

The data used in our experiments are The Saudi Newspapers (SNP) [13], the data set consist of 5121 Arabic documents of different lengths that belongs to 7 categories, the categories are (Culture "الثقافية " , Economics "الإقتصادية" , General "العامة" , Information Technology " تكنولوجيا المعلومات " , Politics "السياسية", Social " الأجتماعية ", Sport " الرياضة"), Table 2 represent the number of documents for each category.

Table 2: Number of Documents per Category

Category Name	Number of Documents
Culture	738
Economics	739
General	728
Information Technology	728
Politics	726
Social	731
Sport	731
Total	**5121**

Arabic text is different from English one since Arabic language is highly inflectional and derivational language which makes monophonically analysis a complex task. Also, in Arabic script, some of the vowels are represented by diacritics which usually left out in the text and it does use capitalization for proper nouns that creates ambiguity in the text [12], [14]. In the Arabic data set we use, each document file was saved in a separate file within the corresponding category's directory.

Moreover, we represented the Arabic data set to a form that is suitable for the classification algorithm. In this phase, we have followed [15], [16], [17] data format and processed the Arabic documents according to the following steps:

1. Each article in the Arabic data set is processed to remove the digits and punctuation marks.

2. We have followed [18] in the normalization of some Arabic letters such as the normalization of (hamza (إ) or (أ)) in all its forms to (alef (ا)).

3. All the non- Arabic texts were filtered.

4. Arabic function words were removed. The Arabic function words (stop words) are the words that are not useful in IR systems e.g. The Arabic prefixes, pronouns, and prepositions.

B. CLASSIFICATION ASSIGNMENT

There are many approaches to assign categories to incoming text such as (SVM) [3], Neural Network [6] and k-nearest neighbor (KNN) [8]. In our paper, we implemented text-to-text comparison (TTC), which is also known as the KNN [8]. KNN is a statistical classification approach, which has been intensively studied in pattern recognition over four decades. KNN has been successfully applied to TC problem, i.e. [19], [8], and showed promising results if compared with other statistical approaches such as Baysian based Network.

IV. Experiment Results

Arabic text is different than English one since Arabic language is highly inflectional and derivational language which makes monophonical analysis a complex task. Also, in Arabic script, some of the vowels are represented by diacritics which usually left out in the text and it does use capitalization for proper nouns that creates ambiguity in the text [14].

Three TC techniques based on vector model similarity (Cosine, Jaccard, and Dice) have been compared in term of F1 measure, which is shown in equation (1). These methods use the same strategy to classify incoming text i.e. KNN. We have several options to construct a text classification method; we compared techniques using IDF term weighting method. All of the experiments were implemented using Java on 3 Pentium IV machine with 1GB RAM.

measures (Recall, Precision, and F1) as the bases of our comparison, where F1 is computed based on the following equation:

$$F1 = \frac{2 * \mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Re}call + \mathrm{Pr}ecision} \qquad (1)$$

Precision and recall are widely used evaluation measures in IR and ML, where according to Table 3 these equation are as following:

$$\mathrm{Pr}\ ecision = \frac{TP}{(TP + FP)} \qquad (2)$$

$$\mathrm{Re}\ call = \frac{TP}{(TP + FN)} \qquad (3)$$

Table 4 gives the F1 results generated by the three algorithms (Cosine, Dice and Jaccard) against seven Arabic data sets; where in each data set we consider 70% of documents arbitrary for training, and 30% for testing. *K* parameter in the KNN algorithm was set to 9.

Table 3. Confusion matrix

Class	Predicted as Actual Class	Predicted as Other
Actual Correct	True Positive (TP)	False Negative (FN)
Other Classes	False Positive (FP)	True Negative (TN)

After analysing Table 4, we found that the Cosine categorize outperformed Dice and Jaccard Algorithms on all measures (F1, Precison and recall).

Particularly, Cosine outperformend Dice and Jaccard on 6,5 data sets respectively with regards to F1 results. Also Recall results obtain that the

Table 4. Results F1, Recall, and Precision of Arabic text categorization

Category Name	Cosine			Dice			Jaccard		
Evaluation measures	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Culture	0.756	0.737	0.746	0.733	0.711	0.722	0.721	0.731	0.726
Economics	0.931	0.927	0.928	0.852	0.887	0.869	0.843	0.83	0.836
General	0.497	0.532	0.514	0.451	0.442	0.446	0.339	0.392	0.364
Information Technology	0.917	0.887	0.902	0.842	0.891	0.865	0.952	0.952	0.952
Politics	0.835	0.873	0.854	0.832	0.921	0.874	0.885	0.842	0.863
Social	0.651	0.623	0.636	0.513	0.52	0.516	0.591	0.542	0.565
Sport	0.917	0.979	0.947	0.96	0.934	0.946	0.911	0.94	0.925

Cosine outperformed Dice and Jaccard on 5,6 data sets respectively. And Precison results obtain that the Cosine also outperformed Dice and Jaccard on 6, 6 data sets respectively.

The average of three measures obtained against seven Arabic data sets indicated that the Cosine dominant Dice and Jaccard.

V. Conclusions and Future Works

In this paper, we investigated different variations of VSM using KNN algorithm, these variations are Cosine coefficient, Dice coefficient and Jacaard coefficient, using IDF term weighting method. The average of three measures obtained against seven Arabic data sets indicated that the Cosine dominant Dice and Jaccard.

We intended to develop new Arabic text classifier to classify Arabic text.

VI. Acknowledgments

The authors are grateful to the World Islamic Sciences and Education University, Amman-Jordan, for supporting this research article.

REFERENCES

[1] F. Sebastiani “Text categorization,” In Alessandro Zanasi (ed.), Text Mining and its Applications, WIT Press, Southampton, UK, 2005, pp. 109—129.

[2] J. Quinlan, C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann, 1993.

[3] T. Joachims “Text Categorisation with Support Vector Machines: Learning with Many Relevant Features,” . Proceedings of the European Conference on Machine Learning (ECML), (pp. 173-142). Berlin, 1998, Springer.

[4] E. D. Wiener, J. O. Perdersen, A. S. Weigend. A Neural Network Approach for Topic Spotting. Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95), 317-332, 1995.

[5] I. Moulinier, G. Raskinis, J. Ganascia, “Text categorization: a symbolic approach” . Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval, 1996.

[6] Sawaf, H. Zaplo,J. and Ney. H. (2001). "Statistical Classification Methods for Arabic News Articles". Arabic Natural Language Processing, Workshop on the ACL'2001. Toulouse, France, July.

[7] T. Tokunaga, M. Iwayama, “Text Categorisation Based on Weighted Inverse Document Frequency”. Department of Computer Science, Tokyo Institute of Technology: Tokyo, Japan, 1994.

[8] Y. Yang. “An evaluation of statistical approaches to text categorization”, Journal of Information Retrieval, 1(1/2):67-88, 1999.

[9] M. Junker, R. Hoch, A. Dengel, “On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy”. in Proceedings of the Fifth International Conference on Document Analysis and Recognition. 1999.

[10] M. Syiam, M. Fayed, Z. T., M. B. Habib, “An Intelligent System For Arabic Text Categorization”, IJICIS, Vol.6, No. 1, 2006.

[11] Laila Khreisat, “Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study”. DMIN 2006: 78-82, 2006.

[12] F. Thabtah, W. Hadi, G. Al-Shammare, “VSMs with K-Nearest Neighbour to Categorise Arabic Text Data.”, In The World Congress on Engineering and Computer Science 2008. (pp.778-781), 22-44 October 2008. San Francisco, USA.

[13] S. Al-Harbi, “Automatic Arabic Text Classification” , JADT’08: 9es Journées internationales d’Analyse statistique des Données Textuelles., pp. 77-83, 2008.

[14] B. Hammo, H. Abu-Salem, S. Lytinen, M. Evens, “QARAB: A Question Answering System to Support the Arabic Language”. Workshop on Computational Approaches to Semitic Languages. ACL 2002, Philadelphia, PA, July. pp. 55-65.

[15] M. Benkhalifa, A. Mouradi, H. Bouyakhf. "Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization," Int. J. Intel Syst (16:8), pp.929-947, 2001.

[16] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer. "An kNN Model-based Approach and its Application in Text Categorization," In proceedings of 5th International Conference on Intelligent Text Processing and Computational Linguistic, CICLing, LNCS 2945, Springer-Verlag, pp.559-570, 2004.

[17] M. El-Kourdi, A. Bensaid, T. Rachidi, “Automatic Arabic Document Categorisation Based on the Naïve Bayes Algorithm”. 20th International Conference on Computational Linguistics . August 28th. Geneva, 2004.

[18] A. Samir, W. Ata, N. Darwish. "A New Technique for Automatic Text Categorization for Arabic Documents," 5th IBIMA Conference (The internet & information technology in modern organizations), 2005, Cairo, Egypt.

[19] Y. Yang, X. Liu, “A re-examination of text categorization methods”, Proceedings of the CAN SIGIR Conference on research and Development in Information Retrieval (SIGIR’99), pp.42-49, 1999.

Dr. jafar Ababneh is an assistant professor. He received his PhD degree from Arab Academy for Banking & Financial Sciences (Jordan) in 2009. He received his M.Sc degree in computer engineering from University of the Yarmouk (Jordan) in 2005. He earned his B.Sc in Telecommunication engineering from University of Mu'ta (Jordan) in 1991.

He joined in 2009 the World Islamic Sciences and Education (WISE) University as a head of the departments of computer information systems and network systems in the school of information technology beside assistance dean of information technology faculty. He has published many research papers, book chapters, and books in different fields of science in refereed journal and international conference proceedings.

His field research lies in development and performance evaluation of multi-queue nodes queuing systems for congestion avoidance at network routers using discrete and continuous time, also his research interests includes computer networks design and architecture, wire and wireless communication, artificial intelligence and expert system, knowledge base systems, security systems, data mining and information.