World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 4, No. 7, 101-104, 2014

Text Summarization as Feature Selection for Arabic Text Classification

Eman Al-Thwaib
Computer Information Systems Department
University of Jordan
Amman, Jordan

Abstract—Text classification (TC) or text categorization task is assigning a document to one or more predefined classes or categories. A common problem in TC is the high number of terms or features in document(s) to be classified (the curse of dimensionality). This problem can be solved by selecting the most important terms. In this study, an automatic text summarization is used for feature selection. Since text summarization is based on identifying the set of sentences that are most important for the overall understanding of document(s). We address the effectiveness of using summarization techniques on text classification.

Another feature selection technique is used, which is Term Frequency (TF) on the same but full-text data set, i.e., before summarization. Support Vector Machine is used to classify our Arabic data set. The classifier performance is evaluated in terms of classification accuracy, precision, recall, and the execution time. Finally, a comparison is held between the results of classifying full documents and summarized documents.

Keywords-Text Categorization; Text Summarization; Support Vector Machine; Feature Selection.

I. INTRODUCTION

With the huge amount of text documents on the web, human needed to save the long time spent dealing with the electronic documents. Automatic text classification and automatic text summarization are two popular solutions proposed to save human time.

Automatic text classification or categorization, as the assignment of a document to a predefined class or category, deals with text documents as features or terms. When dealing with large documents, a problem of huge number of features or terms appears, the need for choosing the most important or informative terms (the so called feature selection) came from here.

Many feature selection techniques were used in many researches (such as TF, TF×IDF, CHI square, etc.). Since 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.

The rest of this paper is organized as follows; related works are reviewed in section 2. Arabic text summarizer used is discussed in section 3. In section 4, the proposed approach is addressed. Text classification and SVM classifier are discussed in section 5. The experimental results are shown in section 6. Finally the paper is concluded in section 7.

II. RELATED WORKS

In [1] the authors applied the combination of word-based frequency and position method on Reuters news corpus to get categorization knowledge from the title field only. Their results indicate that summarization-based categorization can achieve acceptable performance and a very short computation time.

Authors of [2] implemented SVM based text classification system for Arabic language articles. CHI square method was used as a feature selection method in the pre-processing step. F-measure was used to evaluate the classification effectiveness and the result is F=88.11 .

Reference [3] addressed a comparative study of five feature selection methods in text categorization, document frequency (DF), information gain (IG), mutual information (MI), x2-test (CHI), and term strength (TS). The experimental results show that CHI and IG are the most effective when applying the k-Nearest Neighbor (kNN) classifier on the Reuters corpus.

Text summarization was used as a feature selection method in [4]. SVM classifier was applied on the Reuters data set and the short summarized documents were removed. The MI-based feature extraction was used in comparison with seven summarization methods (based on the selection of

sentences with the most important concentration of keywords or title words) and the execution time was calculated. Five of the seven summarization methods outperformed the MI-based feature selection.

Reference [13] addressed text summarization for feature selection as well. kNN and Naïve Bayes (NB) algorithms were used for classifying a data set of 1000 Arabic documents. A comparison was held between classification results of full-text documents (but without using any feature selection technique) and summarized documents. The experimental results showed little bit better accuracy for classifying full documents but shorter execution time and less memory space needed for classifying summarized documents.

Finally, [5] showed that text summarization is a competitive approach for feature selection especially for situations having small training sets. The results were compared to those achieved by the information gain technique. For experiments, a subset of Reuters- 21578 corpus was used. SVM was used as classification algorithm, term frequency as weighting scheme, and the classification accuracy and F1 as evaluation measures.

III. TEXT SUMMARIZATION

Automatic text summarization is the process in which a computer takes a text document(s) as input and produces a summary (or a shortened form) of that document(s) as an output.

Sakhr summarizer [11] is used in this research for summarizing the Arabic documents. It makes it easy to scan just the important sentences within a document by highlighting the most relevant (to the topic of document) sentences within a text. Key words extractor and spelling corrector are used in forming the summary.

IV. THE PROPOSED APPROACH

Figure. 1 shows the full steps of our proposed approach as follows:

1. Documents (after using TF for feature selection) are classified using SVM classifier, so that the class of each document is predicted.
2. The same documents pass a text summarizer, the summaries resulted are classified using SVM, so a class for each document is predicted.
3. The classification results are compared in terms of accuracy, precision, recall, and execution time.

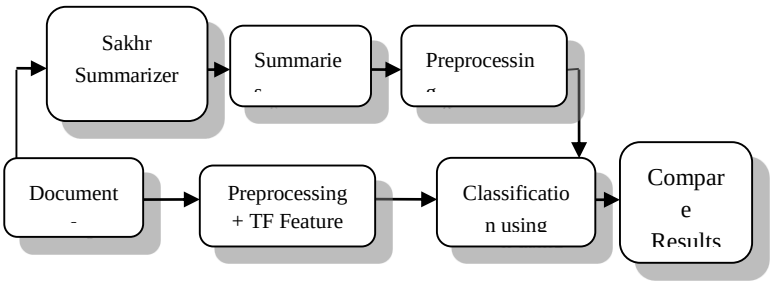

Figure 1. The Proposed Approach

V. TEXT CLASSIFICATION

Text categorization is the problem of automatically assigning text documents to predefined categories/classes. One difficulty of text categorization problems is high dimensionality of the feature space. Feature space can consist of hundreds or thousands of unique terms [12].

In our research, we choose to use the SVM [6],[7],[8],[9],[10] for its high accuracy and an inherent ability to handle large feature spaces such as text [4].

According to [14] a support vector machine constructs a hyperplane or set of hyperplanes in a space, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class as in figure. 2

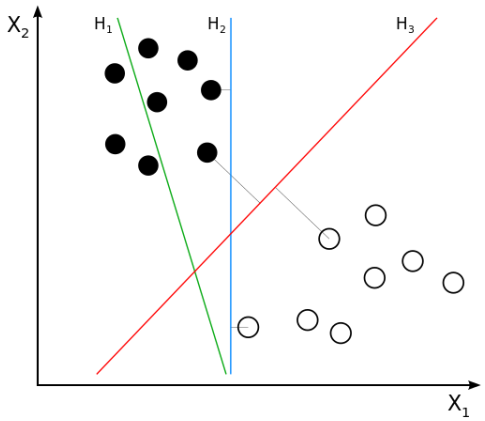

Figure 2. SVM separating hyperplanes

In figure. 2, H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin.

Data Preprocessing:

The data set that we used consists of 800 Arabic text documents. It is a subset of 60913-document corpus collected from many newspapers and other web sites. The 800 documents were pre-classified to four different classes (Economy, Politics, Religion, and Sport), 200 documents for each class.

The Arabic data set documents have been preprocessed according to the way used in [13], each document have been tokenized, i.e. split it into tokens according to the white space position.

Then two copies of the data set are made. From the original or full-text documents, tokens that are less than 3 letters are removed (terms with TF < 3). Sakhr summarization techniques are applied to the second copy of the data set, and a set of 800 summaries is resulted. Then the following preprocessing steps are applied to both copies of data set (full-text and summarized documents):

1. Punctuations (such as ! ؛ , . ؟), symbols (such as < > }]), and digits have been removed. The comma ” ,” has a special case, because it appears sometimes connected to a word

(without a space in between). Our preprocessor searches the beginning and end of tokens for a comma and removes it.

2. Non-Arabic words have been removed.

3. Stop words (such as عن ,لكن,في) have been removed.

4.Remaining terms have been normalized, i.e., Letters " ء", " آ " , " أ " , " ؤ ", " ئ ", and " ئ" have been replaced with "ا", letter " ى " replaced with "ي", and the letter " ة " replaced with " ه ".

VI. RESULTS

As mentioned before, the performance of the SVM classifier (in classifying the full and the summarized documents) is measured with respect to the accuracy, precision, and recall.

Accuracy, precision, recall can be measured by the following equations:

$$\text{Accuracy}=\frac{\text{number of correctly classified documents}}{\text{total number of testing documents}}\times 100\% \quad (1)$$

$$\text{Precision}=\frac{\text{number of correctly classified documents}}{\text{number of all documents assigned to that category by the classifier}} \quad (2)$$

$$\text{Recall}=\frac{\text{number of correctly classified documents}}{\text{number of all documents belonging to that category}} \quad (3)$$

Also the time needed for classification is taken into account.

Table 1 gives the classification accuracy, precision, recall (in average for all categories), and execution time resulted from applying the SVM classifier on the full-text documents and the summarized ones. By analyzing the table, we find that using summarized documents increases the classification accuracy, precision, and recall. But we have needed shorter time for processing the full-text documents because using TF for feature selection resulted in eliminating large number of terms (that have TF less than 3) which resulted in less terms than those in summarized documents. While Sakhr summarizer considers all keywords (like proper names and dates) in forming the summary regardless of number of times it appear.

TABLE 1. THE EXPERIMENTAL RESULTS

Performance Measure	SVM	
	Full-Text Documents	Summarized Documents
Accuracy	83%	94%
Precision	0.821	0.943
Recall	0.809	0.943
Execution Time	4.76 seconds	25.75 seconds

All the experiments were conducted using Weakaito Environment for Knowledge Acquisition (WEKA) [15] where SVM is already implemented in Java.

The data set was tested using k-fold cross-validation method with k=10, where data is divided into 10 equal parts. One part is used for testing and the remaining nine parts are used for training the classifier.

Figure. 3 shows the precision detailed results for classifying full and summarized documents, and figure. 4 shows the detailed results of recall for the four categories (economy, religion, sport, and politics)

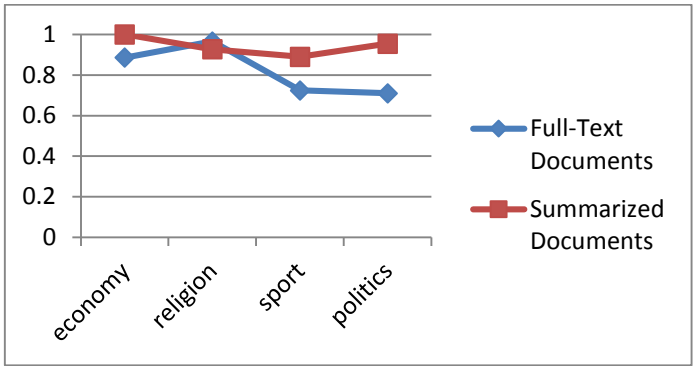

Figure 3. Precision Results

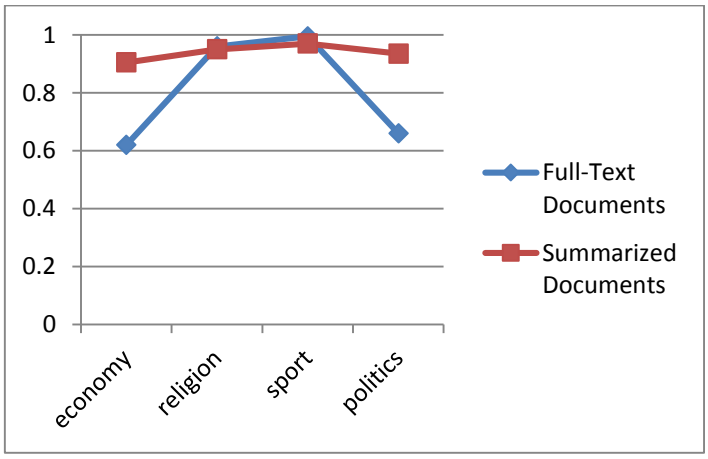

Figure 4. Recall Results

VII. CONCLUSION

In this research, we studied the effect of using automatic text summarization as features selection technique for classifying Arabic documents .We succeed to increase the classification accuracy by using the summarized data set as input for SVM classifier. Our experiments resulted in higher accuracy, precision, and recall but longer execution time for summarized documents (in comparison with full documents).

REFERENCES

[1] Sue J. Ker and Jen-Nan Chen (2000). "A Text Categorization Based on Summarization Technique". In Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 11, pages 79{ 83, Morristown, NJ, USA. Association for Computational Linguistics.

[2] Abdelwadood Moh'd A Mesleh (2007). "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System". Journal of Computer Science 3 (6): 430-435, 2007 ISSN 1549-3636.

[3] Yiming Yang and Jan O. Pedersen (1997). "A Comparative Study on Feature Selection in Text Categorization". Proceedings of ICML-97, pp. 412-420.

[4] Aleksander Kotcz, Vidya Prabakarmurthi, and Jugal Kalita (2001). "Summarization as Feature Selection for Text Categorization". In Proceedings of the 10th International Conference on Information and Knowledge Management, pp. 365-370.

[5] Emmanuel Anguiano-Hernández and Luis Villaseñor Pineda and Manuel Montes-y-Gómez and Paolo Rosso (2010) ."Summarization as Feature Selection for Document Categorization on Small Datasets". IceTAL'10. Pages 39-44.

[6] Saleh Al Saleem (2011). "Automated Arabic Text Categorization Using SVM and NB". International Arab Journal of e-Technology, Vol. 2, No. 2.

[7] Abdelwadood Mesleh (2007)."Support Vector Machines based Arabic Language Text Classification System: Feature Selection Comparative Study". 12th WSEAS Int. Conf. on APPLIED MATHEMATICS, Cairo, Egypt, December 29-31.

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

[9] Thoresten Joachims (1999). "Transductive Inference for Text Classification using Support Vector Machines". Proceedings of the International Conference on Machine Learning (ICML), (pp. 200-209).

[10] Thoresten Joachims (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". In Proceedings of the European Conference on Machine Learning (ECML), pp.173-142, Berlin.

[11] Sakhr company website: http//:www.sakhr.com .last visit in April, 2014.

[12] Peter Bednar, Tomas Futej (2004). "Reduction Techniques for Instance based Text Categorization". In Proceedings of the IFIP TC5/WG 5.5 Sixth IFIP International Conference on Information Technology for Balanced Automation Systems in Manufacturing and Services, Vienna, Austria. ISBN 0-387 22828-4, 475- 480.

[13] Khalil Al-Hindi, Eman Al-Thwaib (2013)."A Comparative Study of Machine Learning Techniques in Classifying Full-Text Arabic Documents versus Summarized Documents". World of Computer Science and Information Technology Journal (WCSIT), Vol. 2, No. 7,pages 126-129.

[14] www.wikipedia.org, the free encyclopedia, last visit june, 2014.

[15] WEKA. Data Mining Software in Java: http://www.cs.waikato.ac.nz/ml/weka. Last visit on May, 2014.