

Text Summarization versus CHI for Feature Selection in Arabic Text Classification

ABSTRACT

Text Classification or Text Categorization (TC) is to assign a document to one or more predefined classes or categories. The need for TC is due to the huge and increasing amount of text documents on the web. The main problem of TC is the high dimensionality of feature space. Many feature selection techniques were used in order to solve this problem. In this research, we consider Arabic text documents and propose text summarization for feature selection. Support Vector Machine (SVM) is then applied to classify the proposed text summarization and the one produced by CHI (one of the well known feature selection techniques). Both classifications are evaluated using precision, recall, accuracy, and execution time. The results for the proposed text summarization demonstrate higher precision, recall, and classification accuracy with negligible higher execution time.

Keywords: Text Classification, Text Summarization, Feature Selection.

1. INTRODUCTION

With the rapid growth of online information, text classification has become one of the key techniques for handling and organizing text documents. The idea of text classification is to categorize textual data into one or more predefined categories. Many classification methods were applied to different datasets, such as kNN, NB, SVM, Rocchio, etc [1] [2] [3] [4] [5] [6].

One of the major problems in text classification that affects its efficiency and increase time complexity is the large number of features or terms of the documents to be classified. Many feature selection techniques have been proposed to reduce this high dimensionality, such as Term Frequency (TF) [1] [7], Document Frequency (DF) [3] [6], CHI [1] [2] [4]] [5] [6] [8], Mutual Information (MI) [2] [3] [4]] [6] , Information Gain (IG), etc [1] [3] [9] [4]] [6] [8].

In this research, we propose automatic text summarization for feature selection. Automatic text summarization is the process in which a computer takes a text document(s) as input and produces a summary of that document(s). A summary must be brief and accurate representation of the important contents of original texts in order to reduce number of features and save time without affecting the classification accuracy. We consider Arabic text documents and propose text summarization for feature selection. Support Vector Machine (SVM) is then applied to classify the proposed text summarization and the one produced by CHI (one of the well known feature selection techniques). This is due to the following:

- Few research works have been conducted on Arabic corpuses. This is mainly because Arabic language is highly rich and requires special treatments [10].
- CHI is one of the well known feature selection techniques [1].
- SVM outperforms many other classification methods in terms of efficiency [4] [5].

The rest of this paper is organized as follows: section 2 briefly discusses related works. The proposed method is addressed in section 3. The experiments are presented in section 4. Finally, a conclusion is presented in section 5.

2. RELATED WORKS

In [1], some feature selection techniques (GHI, IG, TF, and many combinations of them) were used for term space reduction. KNN, NB, Rocchio, and SVM were used to compare results of using the different feature. The comparison showed that combine X^2 statistics with DF or IG to eliminate rare words are the best.

The author of [2] has used five feature selection methods (CHI, NGL, GSS, OR, MI) on an Arabic dataset of 1445 documents. SVM text classifier was applied for full-text documents (i.e. without using FS). A comparison was held among the results of using each one of the feature selection techniques using precision, recall, F1, macro P, and macro R.

A similar research has been conducted in [7], but the comparison was held between Term Frequency (TF) and text summarization.

A comparison of three reduction techniques for reducing feature space (IG, MI, and DF) has been performed in [3]. These techniques have been tested using k-NN classifier. The experiments showed better performance of DF and IG.

Text summarization has been used in [9] versus IG. Using SVM as classification method, it has been showed that summarization performed better for small datasets.

Chi-squared, IG, MI and Symmetrical Uncertainty has been compared and tested in [4] using different classifiers (NB, SVM, decision tree and k-NN). As a result, SVM outperformed other classifiers in all the occasions, NB was the worst in terms of accuracy and IG gives the best results of feature selection.

SVM classification method has been compared to NB and kNN in classifying Arabic Language text in [5] using CHI. The experimental results showed that SVM outperforms NB and kNN.

In [6], DF, IG, MI, Term Strength (TS), and CHI have been compared and tested using kNN and Linear least Squares Fit Mapping (LLSF) classifiers. The experiments showed that CHI and IG outperform other techniques in feature reduction.

Ambiguity Measure (AM) feature selection method with SVM for time complexity reduction has been proposed and compared in [8] to seven different feature selection methods (Odds Ratio, TFICF, TFIDF, Improved Gini Index, IG, Cross Entropy (CE), and CHI). The results showed that AM outperformed the other seven feature selection methods in terms of training time reduction by 50%.

3. THE PROPOSED METHOD

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. In this research, we apply Sakhr [11] summarizer on Arabic documents. This 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 then used in forming the summary. Finally, the highlighted sentences are copied into a new file to form the summarized document. The summary of documents tends to contain the most relevant words and subsequently acts as a feature selection technique. On the other hand, CHI has demonstrated it's self as one of the best feature selection techniques. Therefore, we investigate the use of text summarization and CHI (X^2) as techniques to reduce the number of feature dimensions and hence it is expected to give good classification results. To validate our expectation SVM is used to classify the proposed text summarization and the one produced by CHI. Thus, considering Arabic documents, figure 1 outlines our proposed method where:

1. A sample of Arabic documents that belong to different categories (e.g. politics, economics, etc.) is collected.
2. CHI [15] is used on one copy of the collected documents for feature selection (after implementing preprocessing according to the steps in section 3.1), and the documents are then classified using the SVM
3. Another copy of the original documents is summarized using the summarization technique (SAKHR [11]).
4. The summarized documents are preprocessed as shown in section 3.1.
5. The preprocessed summarized documents are then classified using SVM
6. The results of classification (when using CHI and when using SAKHR summarizer) are compared as in section 4.

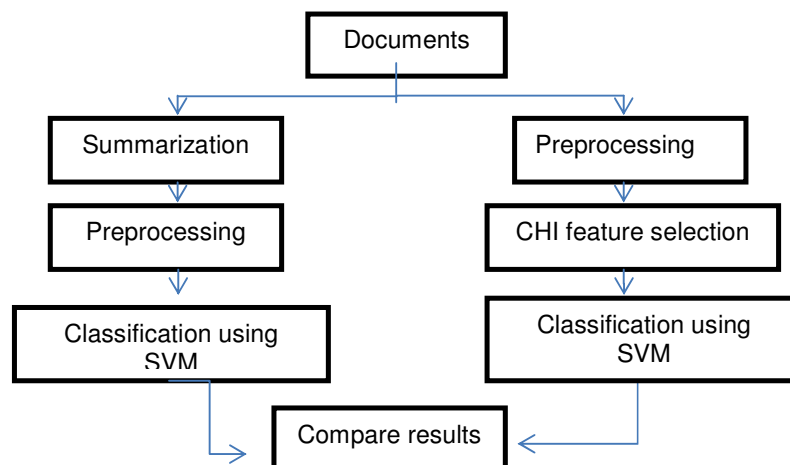


Fig. 1. The Proposed Method

3.1. Arabic Data Preprocessing

The data set/corpus 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 into four different classes (Economy, Politics, Religion, and Sport), 200 documents for each class. The text documents have been preprocessed before being used, each document have been tokenized, i.e. split it into tokens according to the white space position. Tokens that less than 3 letters were removed, then we have followed [13] as follows:

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 frequently occur in all corpus without any added value such as (في, لكن, عن) have been removed .
4. Remaining terms have been normalized, i.e., Letters “ ئ ”, “ ز ”, “ أ ”, “ آ ”, “ ء ”, and “ ي ” have been replaced with “ا”, letter “ى” replaced with “ي”, and the letter “ة” replaced with “ه”.

4. Experimental Results

The proposed method has been applied on the 800 Arabic text documents (section 3.1). For example, figure 2 shows a sample of such documents. Figure 3 and figure 4 show a summarized and a preprocessed version of the considered sample.

The performance of the SVM is measured with respect to the precision, recall, accuracy, and execution time. Recall and precision are defined as in [14] and as given bellow.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

Where TP stands for true positive, FP stands for false positive, and FN stands for false negative.

The SVM classifier that has been applied is the one proposed by Weka Environment for Knowledge Acquisition (WEKA) [12]. Its data set (Arabic documents) has been divided into 70% for training and 30% for testing.

Table 1 shows the classification precision, recall, accuracy, and execution time resulted from applying the SVM classifier on the full-text documents (using CHI as feature selection technique) and the summarized ones.

For further comparison, Table 1 shows the results of a similar work [7] that has been conducted using Term Frequency (TF). The obtained results as shown by Table 1 demonstrate a fact that the text

153 summarization for feature selection increases precision, recall, and accuracy of the classification as
154 compared to CHI and TF, with negligible difference in execution time (25.75 seconds versus 4.97 and
155 4.76 seconds respectively).
156 Furthermore, a noticeable reduction in the size of the original document is obtained. The size of the
157 original document is reduced by 70% as shown by figure1 and figure 3.

وزير المالية السعودي: مقر البنك المركزي الخليجي لن يطرح للتفاوض من جديد

مسقط: رويترز- داليا مرزبان

قال ابراهيم العساف وزير المالية السعودي أن بلاده وثلاث دول خليجية أخرى ستمضي في خطة الوحدة النقدية وان مقر البنك المركزي الخليجي لن يطرح للتفاوض من جديد. وقال العساف في مقابلة مع رويترز أمس في سلطنة عمان بعد أقل من أسبوعين من انسحاب الإمارات العربية المتحدة من الخطة: لن تخرج الخطة عن مسارها بل ستستمر. ستمضي الوحدة النقدية كما هو مخطط لها". وأضاف "مادمننا نمضي في الاتجاه الصحيح فهذا أهم شيء.

وكانت الإمارات ثاني أكبر اقتصاد في العالم العربي قد انسحبت الشهر الماضي من خطة لإصدار عملة موحدة احتجاجا على قرار اختيار الرياض مقرا للبنك المركزي المشترك. وتشارك في خطة الوحدة النقدية إلى جانب السعودية كل من الكويت وقطر والبحرين

"وردا على سؤال عما إذا كان مقر البنك المركزي مطروحا لإعادة التفاوض قال العساف "لا هناك قرار اتخذناه قادتنا وكان وزير الخارجية الإماراتي صرح لرويترز في وقت سابق من الشهر أن بلاده ستدرس إعادة الانضمام إلى الوحدة النقدية إذا تغيرت الشروط ووافق جيرانها على السماح بأن تكون الإمارات مقرا للبنك المركزي.

وقال الوزير الشيخ عبد الله بن زايد آل نهيان أن الاقتصاد المفتوح الذي تتمتع به الإمارات هو الأكثر ملائمة في منطقة الخليج لاستضافة البنك المركزي.

وفي وقت سابق من يوم السبت قال الأمين العام لمجلس التعاون الخليجي أن من المقرر أن تجتمع الدول الأربعة الأخرى المشاركة يوم السابع من يونيو للتوقيع على اتفاق الوحدة النقدية. ويتساءل المحللون عما إذا كان انسحاب الإمارات ثاني أكبر السعودية يمكن أن يخرج المشروع الذي يواجه مشكلات منذ فترة عن مساره. وقال العساف أن من بين اقتصاد عربي بعد المزايا الرئيسية للعملة الموحدة خفض تكاليف المعاملات بين الدول الأعضاء في مجالات مثل التجارة والسياحة. وأردف "لديك تكاليف المعاملات التجارية كما أن لديك مخاطر تباين أسعار الصرف. حتى في ظل ربط معظم العملات حاليا بالدولار لن يكون ذلك هو الحال بالضرورة في المستقبل.

العملة الموحدة ستقلل المخاطر وسيؤثر هذا بشدة على قرارات الاستثمار والإيداع وأي نوع من المعاملات التجارية". وتابع أن "العملة الموحدة ستكون أيضا أكبر منطقة مصدرة للنفط في العالم من تكوين "كتلة نقدية رئيسية.

Fig. 2. A sample Document

وزير المالية السعودي: مقر البنك المركزي الخليجي لن يطرح للتفاوض من جديد

مسقط: رويترز- داليا مرزبان

قال ابراهيم العساف وزير المالية السعودي أن بلاده وثلاث دول خليجية أخرى ستمضي في خطة الوحدة النقدية وان مقر البنك المركزي الخليجي لن يطرح للتفاوض من جديد.

وكانت الإمارات ثاني أكبر اقتصاد في العالم العربي قد انسحبت الشهر الماضي من خطة لإصدار عملة موحدة احتجاجا على قرار اختيار الرياض مقرا للبنك المركزي المشترك.

وكان وزير الخارجية الإماراتي صرح لرويترز في وقت سابق من الشهر أن بلاده ستدرس إعادة الانضمام إلى الوحدة النقدية إذا تغيرت الشروط ووافق جيرانها على السماح بأن تكون الإمارات مقرا للبنك المركزي

وقال الوزير الشيخ عبد الله بن زايد آل نهيان أن الاقتصاد المفتوح الذي تتمتع به الإمارات هو الأكثر ملائمة في منطقة الخليج لاستضافة البنك المركزي.

Fig. 3. A Summarized Document

وزير المالية السعودي مقر البنك المركزي الخليجي يطرح للتفاوض جديد مسقط رويترز داليا مرزبان ابراهيم العساف وزير
المالية السعودي بلاده وثلاث دول خليجية ستمضي خطه الوحدة النقدية مقر البنك المركزي الخليجي يطرح للتفاوض جديد
الامارات ثاني اكبر اقتصاد العالم العربي انسحبت الشهر الماضي خطه لاصدار عمله موحده احتجاجا قرار اختيار الرياض
مقرا للبنك المركزي المشترك وزير الخارجية الاماراتي صرح لرويترز وقت سابق الشهر بلاده ستدرس اعاده الانضمام
الوحدة النقدية تغيرت الشروط ووافق جيرانها السماح الامارات مقرا للبنك المركزي للوزير الشيخ عبد الله زايد نهيان
الاقتصاد المفتوح تتمتع الامارات ملامه منطقه الخليج لاستضافه البنك المركزي.

Fig. 3. A Preprocessed Document

Table 1: The Experimental Results

SVM	PERFORMANCE MEASURE			
	PRECISION	RECALL	ACCURACY	EXECUTION TIME
FULL-TEXT DOCUMENTS WITH CHI	0.825	0.808	80.8%	4.97 SECONDS
FULL-TEXT DOCUMENTS WITH TF	0.821	0.809	83%	4.76 SECONDS
SUMMARIZED DOCUMENTS	0.947	0.94	94%	25.75 SECONDS

5. CONCLUSION

This research aims to study the effect of using text summarization as a feature selection tool on the text classification. To do so, CHI is used as a comparative feature selection tool. Then, a classification is performed using SVM for summarized Arabic text documents and for their respective original ones as processed by CHI. The classification efficiency was measured in terms of precision, recall, accuracy, and execution time. We conclude that text summarization increases text classification efficiency but with a negligible longer execution time. However, it is still in seconds.

REFERENCES

- [1] Rogati M, Yiming Y. High-performing feature selection for text classification. ACM 1-58113-492-4/02/0011.2002
- [2] Mesleh A. Support vector machines based arabic language text classification system: feature sfelection comparative study. Advances in Computer and Information Sciences and Engineering. 2008:11–16
- [3] Bednar P. Futej M. Reduction techniques for instance based text categorization. Machine Learning journal. 2000 ; 32: 257-286.
- [4] Subhajit S, Saptarsi, G. Empirical study on filter based feature selection methods for text classification.. International Journal of Computer Applications. 2013; 81(6).
- [5] Mesleh A. Chi square feature extraction based svms arabic language text categorization system. Journal of Computer Science. 2007; 3 (6): 430-435.
- [6] Yimig Y, Bederson, J. A comparative study on feature selection in text categorization. In: J. D. H. Fisher, editor, The Fourteenth International Conference on Machine Learning. Morgan Kaufmann 1997; 412-420.

- 196 [7] Al-Thawib, E. A Text summarization as feature Selection for Arabic text Classification. The World of Computer
197 Science and Information Technology Journal (WSCIT). 2014; 4(7): 101.104.
- 198 [8] Mengle, S, Goharian, N. Using ambiguity measure feature selection algorithm for support vector machine
199 Classifier. Proceedings of the 2008 ACM symposium on Applied computing. 2008; 916-920.
- 200 [9] Anguiano-Hernández, E , Villaseñor-Pineda, L, Montes-y-Gómez, M, Rosso P. Summarization as feature selection
201 for document Categorization on Small Datasets. Notes in Computer Science. LNAI 6233. Springer-Verlag Berlin
202 Heidelberg. 2010;. 39–44.
- 203 [10] Anwar, A, Salama, G, Abdelhalim, M. Video classification and retrieval using arabic closed caption". 2013,
204 Accessed May 2016. Available: <http://icit.zuj.edu.fo/icit13/Papers%20list>.
- 205 [11] Sakhr company website. 2016. Accessed May 2016. Available: <http://www.sakhr.com>
- 206 [12] Khorsheed, M, Al-Thubaity, A. Comparative evaluation of text classification techniques using a large diverse
207 Arabic dataset. Language Resources and Evaluation. 2013;47(2): 513-538..
- 208 [13] Samir, A, Ata, W, Darwish, N. New technique for automatic text categorization for arabic documents. Proceedings
209 of the 5th Conference of the Internet and Information Technology in Modern Organizations, Cairo, Egypt. 2005; 13-
210 15
- 211 [14] Ikonomakis, M, Kotsiantis, S, Tampakas, V. Text classification using machine Learning techniques. Wseas
212 Transactions on Computers. 2005; 4(8): 966-974.
- 213 [15] WEKA. Data Mining Software in Java. 2015. Accessed May 2015. Available: <http://www.cs.waikato.ac.nz/ml/weka>.
- 214