Chapter 12
An Experimental Study for the Effect of Stop Words Elimination for Arabic Text Classification Algorithms

Bassam Al-Shargabi
Isra University, Jordan

Fekry Olayah
Isra University, Jordan

Waseem AL Romimah
University of Science and Technology, Yemen

ABSTRACT

In this paper, an experimental study was conducted on three techniques for Arabic text classification. These techniques are Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Naïve Bayesian (NB), and J48. The paper assesses the accuracy for each classifier and determines which classifier is more accurate for Arabic text classification based on stop words elimination. The accuracy for each classifier is measured by Percentage split method (holdout), and K-fold cross validation methods, along with the time needed to classify Arabic text. The results show that the SMO classifier achieves the highest accuracy and the lowest error rate, and shows that the time needed to build the SMO model is much lower compared to other classification techniques.

INTRODUCTION

Text classification (TC) is the process of classifying documents into a predefined set of categories based on the content of documents. Arabic is a greatly inflectional and derivational language which makes text and web mining a difficult task. The organization of text in categories allow the user to limit the target of a search submitted to information retrieval systems, to explore the collection and to find relevant information they need with poor knowledge about the keywords

DOI: 10.4018/978-1-4666-2157-2.ch012

of a theme (Al-Harbi, Almuhareb, Al-Thubaity, Khorsheed, & Al-Rajeh, 2008).

The aim of the classification techniques is to minimize information loss while maximizing reduction in dimensionality. The set of documents to be categorized must transformed into a set of feature vectors in a relatively low dimensional feature space. The set of reduced feature vectors is then fed to the text classifier as input. The SMO and NB classifier must be trained before it can be used for text categorization. The neural network is used to train classifier using the back propagation learning rule based on supervised learning. A set of training documents along with a set of pre-defined categories that documents belong to are required.

However, in Arabic language there is a very limited work in automatic classification. Classifying Arabic text is different than classifying English language because Arabic is a highly inflectional and derivational language which makes monophonical analysis a very complex task (Al-Shalabi, Kanaan, Jaam, Hasnah, & Hilat, 2004; Sawaf, Zaplo, & Ney, 2001).

Few researches tackled the area of Arabic text classification. A statistical method called maximum entropy is used to classify Arabic News articles (El Kourdi, Bensaid, & Rachidi, 2004). Another statistical classifier based on NB used to categorize Arabic web documents using the root based stemmer to extract the roots of the words (Al-Shalabi, Kanaan, Jaam, Hasnah, & Hilat, 2004). Although in Gharib and Badieh (2009) they used SVM to classify Arabic text and compared result with K-Nearest Neighbor classifier.

In this paper we compared the accuracy of three classifiers along with studying the effects of elimination of stop word for Arabic text. The first classifier is support vector machine (SMO) with sequential minimal optimization (SMO), Naïve Bayesian (NV), and J48.A standard Arabic data set was used to test the above techniques.

The rest of the paper organized as follows: First we discuss preprocessing of Arabic text document and introduce proposed comparative process for the three classifying techniques used in this paper. Next, we present Experimental results and discussions, and finally conclusion is presented.

PREPROCESSING AND CLASSIFICATION PROCESS

The preprocessing of data set deals with the elimination of non-meaningful words which do not indicate semantic content of the document. Some words appear in the sentences and don't have any meaning or indications about the content such as (so لذلك, with بالإشارة, confirmation تأكيدا, for بالنسبة) or appearing frequently in the document like pronouns such as (he هو, she هي, they هم). Although the prepositions like (from من, to إلى, in في, about عن) or demonstratives like (this هذا, these هؤلاء, there أولئك) or interrogatives like (where أين, which أي, who من). These words may have a bad effect on statistical information and co-occurrence of the words as stated in Abo Alkhair (2006) and Said, Wanas, Darwish, and Hegazy (2009).

Also the numbers and symbols like (@, #, &, %, *) and some words that indicates a sequence of the sentences like (firstly أولا, secondly ثانيا, thirdly ثالثا), these words will be considered as an Arabic stop words. Some Arabic documents may contain foreign words, especially science documents, these words are alo considered as stop words as in Al-Shalabi, Kanaan, Jaam, Hasnah, and Hilat (2004) and El-Kourdi, Bensaid and Rachidi (2004).

In fact, a word which occurs in 80% of the documents in the collection is useless for purposes of retrieval. Such words are frequently referred to as stop words and are normally filtered out as potential index terms. Articles, prepositions, and conjunctions are natural candidates for a list of stop words. Elimination of stop words has an additional important benefit. It reduces the size of the indexing structure considerably. In fact, it is typical to obtain a compression in the size