The Effect of Stemming on Arabic Text Classification: An Empirical Study

Abdullah Wahbeh, Dakota State University, USA
Mohammed Al-Kabi, Yarmouk University, Jordan
Qasem Al-Radaideh, Yarmouk University, Jordan
Emad Al-Shawakfa, Yarmouk University, Jordan
Izzat Alsmadi, Yarmouk University, Jordan

ABSTRACT

The information world is rich of documents in different formats or applications, such as databases, digital libraries, and the Web. Text classification is used for aiding search functionality offered by search engines and information retrieval systems to deal with the large number of documents on the web. Many research papers, conducted within the field of text classification, were applied to English, Dutch, Chinese, and other languages, whereas fewer were applied to Arabic language. This paper addresses the issue of automatic classification or classification of Arabic text documents. It applies text classification to Arabic language text documents using stemming as part of the preprocessing steps. Results have showed that applying text classification without using stemming; the support vector machine (SVM) classifier has achieved the highest classification accuracy using the two test modes with 87.79% and 88.54%. On the other hand, stemming has negatively affected the accuracy, where the SVM accuracy using the two test modes dropped down to 84.49% and 86.35%.

Keywords: Arabic Text Classification, Decision Tree, Naïve Bayes Classifier (NB), Natural Language Processing, Stemming, Support Vector Machine (SVM), Text Classification

1. INTRODUCTION

The tremendous growth of available Arabic text documents on the Web and databases have posed a major challenge on researchers to find better ways to deal with such huge amount of information in order to enable search engines and information retrieval systems to provide relevant information accurately, which has become a crucial task to satisfy the needs of different end users.

Text classifications, and its techniques, have become a major tool for dealing with the large amount of available data on the Web and databases. Text classification is the task of automatically assigning text documents to one or more predefined categories based on content and linguistic features (Gharib et al., 2009;
Mesleh et al., 2007; Rahman et al., 2003; Zubi, 2009; Al-Harbi et al., 2008). Several researches applied text classification and its techniques to English and other European languages. On the other hand, few researchers have addressed the issue of Arabic text classification.

Text preprocessing and preparation; especially for Arabic, is a crucial task in several applications including: information retrieval, text mining, and natural language processing where the processing tasks include different stages such as: stop word removal and stemming. Stemming tries to reduce a word to its stem (Al-Shammari et al., 2008), stemming process uses word morphological analysis in order to get the word’s stems (Sembok et al., 2011).

Stemming is a very important technique that is usually used in information retrieval and data mining as well as many other NLP applications. Stemming is important for some natural languages and unimportant in others. As reported by Sembok et al. (2011) and Al-Shammari (2008), stemming has the following benefits:

- Stemming helps in reducing the size of the index terms.
- Stemming is used in information retrieval systems to reduce variant word forms to common roots in order to improve retrieval effectiveness.

Arabic is a language used by millions of people around the world in more than 25 countries. Arabic consists of 28 letters, three vowels and the remaining letters are consonants. Each letter has a different style depending on its position in the word (Duwari, 2007; Kadri et al., 2006) Arabic is a highly inflectional and derivative language which makes morphological analysis a very complex task. Moreover, Arabic do not use capitalizations in order to differentiate nouns form other words in documents (El-Halees, 2010). Arabic words have two distinctive genders, feminine and masculine; three numbers, singular, dual, and plural; and three grammatical cases, nominative, accusative, and genitive (Omer et al., 2010). Finally, the Arabic language consists of three types of words; nouns, verbs and particles; where nouns and verbs are derived from a limited set of about 10,000 roots (Said et al., 2009). All these characteristics make Arabic text classification a difficult task when comparing it with other text classification tasks that deal with English and other languages.

Many text classification techniques used in data mining and machine learning exist in literature such as: Support Vector Machine (SVM) (Gharib et al., 2009; Mesleh et al., 2007; Zhuang et al., 2005), Naïve Bayes (NB) (Gharib et al., 2009; Zubi, 2009; Duwari, 2007; El-Kourdi et al., 2004), K-Nearest Neighbor (KNN) (Zubi, 2009; Duwari, 2007; Zhou et al., 2003), Rocchio (Gharib et al., 2009), and Decision Tree (Al-Harbi et al., 2008). However, in this work we addressed three of those classifiers namely, Support Vector Machine (SVM), Naïve Bayes, and Decision Tree.

The main aim of the paper is to determine the effect of applying stemming as a preprocessing step on the classification performance of the three selected classifiers. All tests are done with the idea that stemming improves the classification accuracy for the classifiers under test as well as reducing the time required for building the model for testing purposes.

1.1. Research Questions

The problem statement or the research questions that this paper is trying to tackle are largely two questions:

- What is the impact of stemming on the text classification process? Will the stemming process always improve the accuracy of classification?
- What is the best text classification algorithm that can be used particularly for Arabic language classification?

The rest of the paper is organized as follows. Section 2 gives related works in the field of text classification. Section 3 demonstrates the methodology used to achieve the required goals.
www.igi-global.com/chapter/information-grounds-theory-1999-2004/127129?camid=4v1a