A Study of Text Preprocessing Tools for Arabic Text Categorization

Dina A. Said[1]; Nayer M. Wanas[1]; Nevin M. Darwish[2]; Nadia H. Hegazy[1]

[1]Pattern Recognition and Information Systems Group
Informatics Department
Electronics Research Institute, Cairo, Egypt

[2]Department of Computer Engineering
Faculty of Engineering
Cairo University, Cairo, Egypt

dasaid@ucalgary.ca, nwanas@mcit.gov.eg, ndarwish@ieee.org, nhegazy@mcit.gov.eg

Abstract

Text preprocessing is an essential stage in text categorization (TC) particularly and text mining generally. Morphological tools can be used in text preprocessing to reduce multiple forms of the word to one form. There has been a debate among researchers about the benefits of using morphological tools in TC. Studies in the English language illustrated that performing stemming during the preprocessing stage degrades the performance slightly. However, they have a great impact on reducing the memory requirement and storage resources needed. The effect of the preprocessing tools on Arabic text categorization is an area of research. This work provides an evaluation study of several morphological tools for Arabic Text Categorization. The study includes using the raw text, the stemmed text, and the root text. The stemmed and root text are obtained using two different preprocessing tools. The results illustrated that using light stemmer combined with a good performing feature selection method enhances the performance of Arabic Text Categorization especially for small threshold values.

1. Introduction

Text Categorization (TC) is the process of assigning a given text to one or more categories. This process is considered as a supervised classification technique, since a set of labeled (pre-classified) documents is provided as a training set. The goal of TC is to assign a label to a new, unseen, document (Sebastiani, 2002).

TC can play an important role in a wide variety of areas such as information retrieval, news recommendation, word sense disambiguation, topic detection and tracking, web pages classification, as well as any application requiring document organization. There has been a debate among researchers about the benefits of using morphological tools in TC. Studies in the English language illustrated that performing stemming during the preprocessing step degrades the performance slightly (Silvatt and Ribeirot, 2003; Song et al., 2005). The experiment conducted by (Debole and Sebastiani, 2005) illustrates that selecting 10% of features exhibits the same classification performance as when using all the features when using SVM in classification. This may indicate that using preprocessing tools and dimensionality reduction techniques is not necessary, for the English language, from the performance point of view when using a robust classifier such as SVM. However, preprocessing tools are essential for decreasing the training time and storage required as indicated by (Zhu et al., 2005).

The main objective of this work is to evaluate preprocessing tools w.r.t. the Arabic language in TC. The reason of choosing the Arabic language is that it is highly derivative where tens or even hundreds of words could be formed using only one root. Furthermore, a single word may be derived from multiple roots (Attia, 2000). Unlike the English language, there are two main approaches to perform Arabic text preprocessing; (i) the stem-based approach, and (ii) the root-based approach. In the stem-based approach, prefixes, and suffixes are removed from the word to extract the word stem. This stem may be further processed to compass the word root in the root-based approach (Darwish, 2003). As an example, the stem of "Ketabhom (their book)" is "Ktab (book)" and its root is "Ktb (wrote)" while the stem of the word "Ktateeb (places for learning Quran)" is the same word "Ktateeb" but the root is "Ktb (wrote)".

Unfortunately, the research in the area of Arabic preprocessing tools is fairly limited. Early studies performed on IR indicated that using root words is better than using stemmed words as mentioned by (Darwish, 2003). Other studies, by (Larkey et al., 2002; Moukdad, 2006), reported that the stem-based approach is superior to the root-based approach. An experiment performed by (Darwish et al., 2005) showed that using context to improve the root extraction process may enhance the process of IR slightly compared to the stem-based approach. However, the context root extraction is computationally expensive compared with the stemming. On the contrary, Brants et al. (Brants et al., 2002) reported that performing stemming to the Arabic text increases the ambiguity, and hence using the raw text may be better.

Due to this contradiction, the main goal of this study is to compare different preprocessing tools for Arabic TC to evaluate their performance. The experiments have been conducted using the raw text, the stemmed text, and the root text. The stemming and root extraction have been performed using two different set of algorithms. The first set was implemented by Kareem Darwish (Darwish, 2002; Darwish, 2003) which consists of Al-Stem stem-based (AS) system and Sebawai root-based (SR) system. The second set was provided by (Attia, 2000), which includes RDIMORPHO3 stemmer (MS) and RDI MORPHO3 root extractor (MR).

2. Arabic Text preprocessing

The Arabic language consists of three types of words; nouns, verbs and particles. Nouns and verbs are derived from a limited set of about 10,000 roots (Darwish, 2002). Templates are applied to the roots in order to derive nouns and verbs by removing letters, adding letters, or including

infixes. Furthermore, a stem may accept prefixes and/or suffixes in order to form the word (Darwish, 2003). In the following we will provide a brief description to the morphological tools used in this study.

2.1 RDI MORPHO3

This system uses rules in conjunction with statistics in order to build a list of possible prefix-suffix template combinations **(**Attia, 2000). These combinations are used in order to transform the word to a root. The main disadvantage of this system is that the rules are built manually which is time consuming and demanding a deep knowledge of the Arabic language. The output of MORPHO3 system is a morphological analysis of the words including its root, stem, meaning of prefixes and suffix, etc…

2.2 Sebawai root extractor (SR)

Sebawai is very similar to MORPHO3 root extractor. However, it uses automatic rules rather than manual rules (Darwish, 2003). Rules have been obtained through training the system with a list of word-root pairs. The author suggests obtaining the training list by three ways; (a) manual construction, (b) using another morphological analyzer tool such as MORPHO3, or (c) parsing a dictionary.

2.3 Al-Stem Stemmer (AS)

Al-Stem is considered a light stemmer where a predefined list of prefixes and suffixes is removed if they are found at the beginning or the end of the word (Darwish, 2003). Darwish obtained this list from the training stage of Sebawai. The list is then purified by examining it manually. Light stemmers have been used by (Larkey et al., 2002) in Arabic information retrieval and have showed better performance than using the root and raw texts.

3. Experimental Setup

3.1 Stop word Removal

In this research, stop words removal has not been applied in the preprocessing of the Arabic documents. The diversity of stop word list depends on the preprocessing tool used. The main objective of these experiments is comparing different pre-processing tools. Therefore, stop words have not been removed to isolate their effect and conduct fair experiments.

3.2 Feature Scoring Methods

As mentioned previously, feature selection has a great impact on reducing the training time and storage required. Additionally, different feature selection methods may affect the performance when selecting features that are less than 10% of the total number of features. In this study, we evaluate the performance of the preprocessing tools using the following feature scoring methods. These methods have been widely used, and have shown to be among the top performing methods in TC (Yang and Pedersen, 1997; Sebastiani, 2002).

• **Document Frequency (DF)** assumes that the single occurrence of the word in a document has the same importance as its multiple occurrences. DF is calculated by counting the number of documents where a specific word, wk, occurs (Yang and Pedersen, 1997).

• **Information Gain (IG***)* is the number of bits gained, for a certain category, by knowing the presence or absence of a word in the document (Sebastiani, 2002). IG is defined as:

$$IG(w_k, c_i) = \sum_{c \in \{c_i, \overline{c_i}\}} \sum_{w \in \{w_k, \overline{w_k}\}} p(c|w) log \frac{p(c|w)}{p(w)p(c)} \qquad (1)$$

• ***Mutual Information (MI)*** measures the mutual dependency between the word w_k and the category c_i. MI is presented by equation 2 according to (Bekkerman et al., 2003).

$$MI(w_k, c_i) = p(w_i|c_i) log \frac{A(w_k, c_i) \times N(Tr)}{N(c_i) \times N(w_k)}. \qquad (2)$$

where $N(Tr)$ is the number of documents in the training set, $A(w_k, c_i)$ is the number of times a word w_k and a category c_i co-occur, $N(c_i)$ is the number of documents in category c_i, and $N(w_k)$ is the number of documents in which feature w_k occurs.

• ***Correlation Coefficient (CC)*** is the square root of the chi square (χ^2) feature scoring method. The χ^2 measures the lack of independency between a word, w_k, and a category, c_i (Yang and Pedersen, 1997).

3.3 Datasets

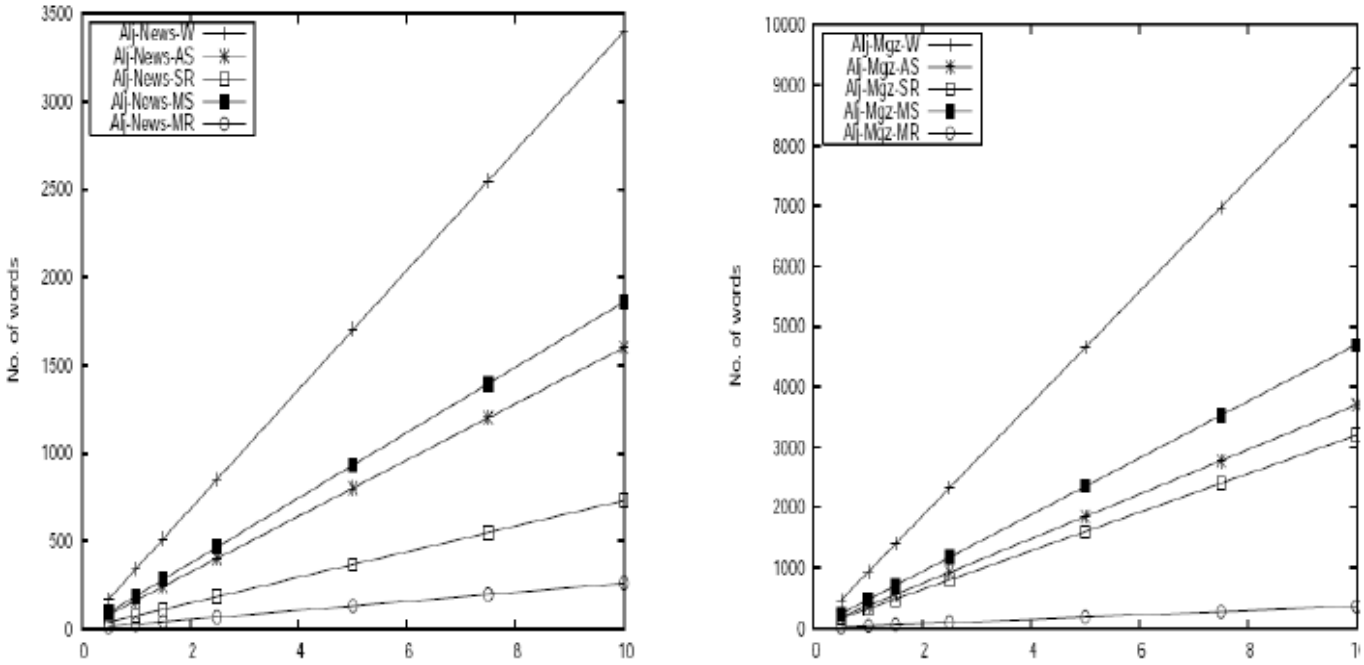
Two datasets are used in this study which are:

– **Aljazeera News Arabic Dataset (Alj-News)**[1] is a collection of 1500 Arabic news documents obtained from Aljazeera online news agency[2] (Mohamed et al., 2005). These documents are evenly distributed among five categories. It is worth noting that the number of documents in this dataset is small and the diversity among the nature of categories is big. This significantly simplifies the classification process. In order to compare the results of this work with the results of (Mohamed et al., 2005), the same split of training and testing files was adopted.

– **Al-jazirah Magazine Arabic Dataset (Alj-Mgz)**[3] is an Arabic dataset collected manually from Al-jazirah online

[1] Available online at http://filebox.vt.edu/users/dsaid/Alj-News.tar.gz

[2] http://www.aljazeera.net/



(a) Alj-News (b) Alj-Mgz

Figure 1: Vocabulary size at different threshold values for (a) Alj-News; and (b) Alj-Mgz datasets using different threshold values (0.5% – 10%) of unique words in the dataset where W is using raw word, AS is Al-Stem stemmer, SR is Sebawai root extractor, MS is RDI MORPHO3 stemmer, and MR is RDI MORPHO3 root extractor.

newspaper[4]. The dataset consists of 4470 articles published from 2001 to 2005. Since there is no standard split to the training and test document of this dataset, cross validation is performed. Five random chosen splits were constructed such that the training documents in each split represent four fifth of the total number of documents. The results of experiments conducted on Alj-Mgz show both the mean and standard deviation using the five different splits.

3.4 Performance Evaluation

MicroF$_1$ and *MacroF*$_1$ tests (Lewis and Ringuette, 1994) are adopted for effectiveness evaluation. They are based on F_1 which combines recall and precision in an equally weighted manner. Equation 3 shows how F_1 is calculated.

$$F_1 = \frac{2 * recall * precision}{recall + precision} \quad (3)$$

The *MacroF*$_1$ test is simply the average of the F_1 of all categories while the *MicroF*$_1$ test calculates recall and precision of the whole dataset and then finds the *F1*. The *MacroF*$_1$ test equally weights all categories, and thus it is influenced by the performance of rare categories. However, the *MicroF*$_1$ test equally weights all the documents, and therefore it is affected by the performance of frequent categories (Sebastiani, 2002). Since the datasets used in this experiment are not highly skewed, only the results of *MicroF*$_1$ have been reported for the sake of space limitation.

3.5 Classification

Support Vector Machine (SVM) has been shown to be among the best performing classifiers in TC applications (Debole and Sebastiani, 2005). In this study, we apply classification using the SVM-light[5] (Joachims, 1999). A linear kernel function has been used in these experiments and the kernel parameters have been initialized with the default SVM-light values. As a future work, we would like to investigate the effect of changing the kernel function as well as other kernel parameters.

[3] Available online at http://filebox.vt.edu/users/dsaid/Alg-Mgz.tar.gz

[4] http://www.al-jazirah.com/

[5] SVM-light is publicly published at http://svmlight.joachims.org.

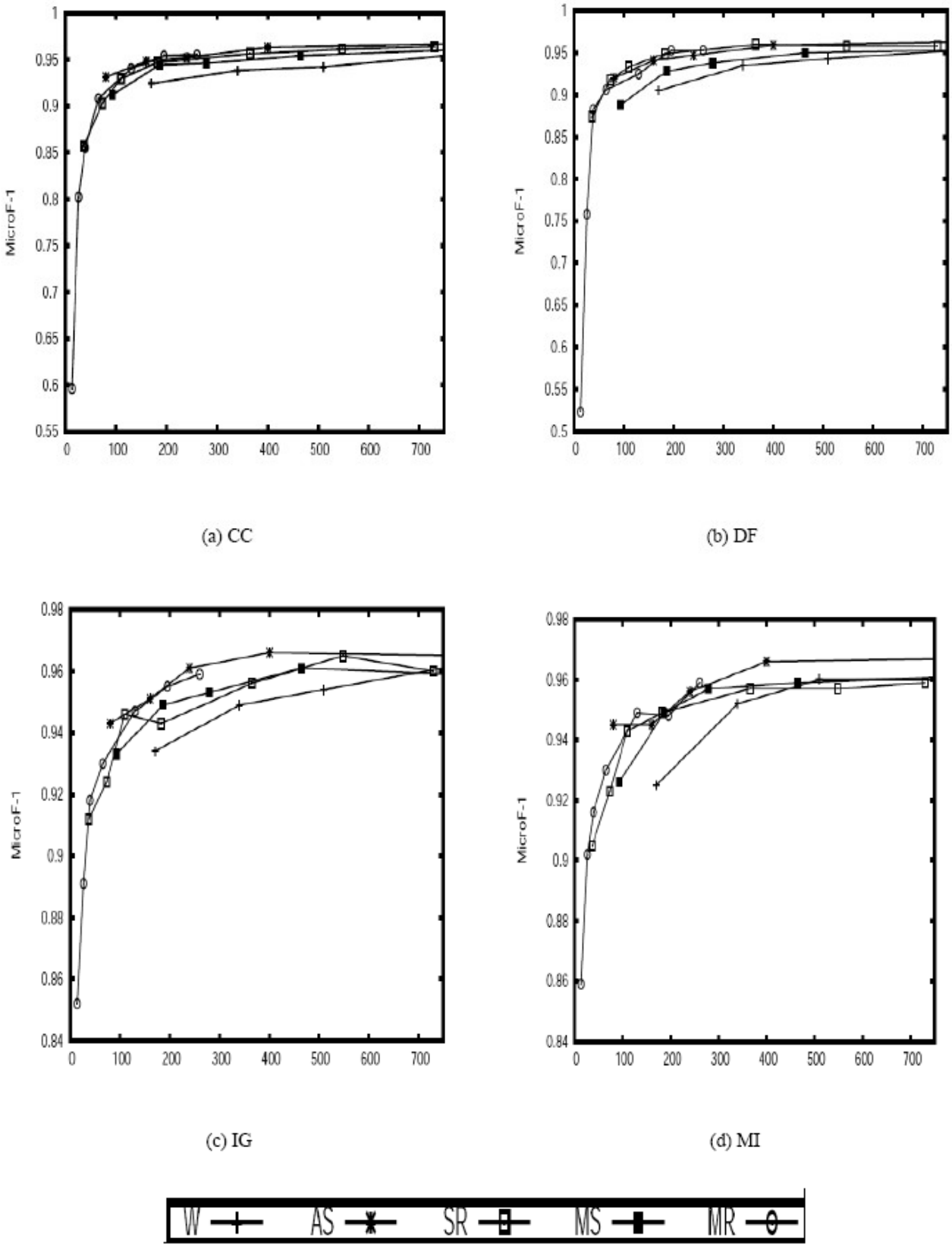


Figure 2 *MicroF*$_1$ of Alj-News dataset using (a) CC; (b) DF; (c) IG; and (d) MI feature scoring methods with different threshold values (0.5% – 10%) of unique words in the dataset where W is using raw word, AS is Al-Stem stemmer, SR is Sebawai root extractor, MS is RDI MORPHO3 stemmer, and MR is RDI MORPHO3 root extractor.

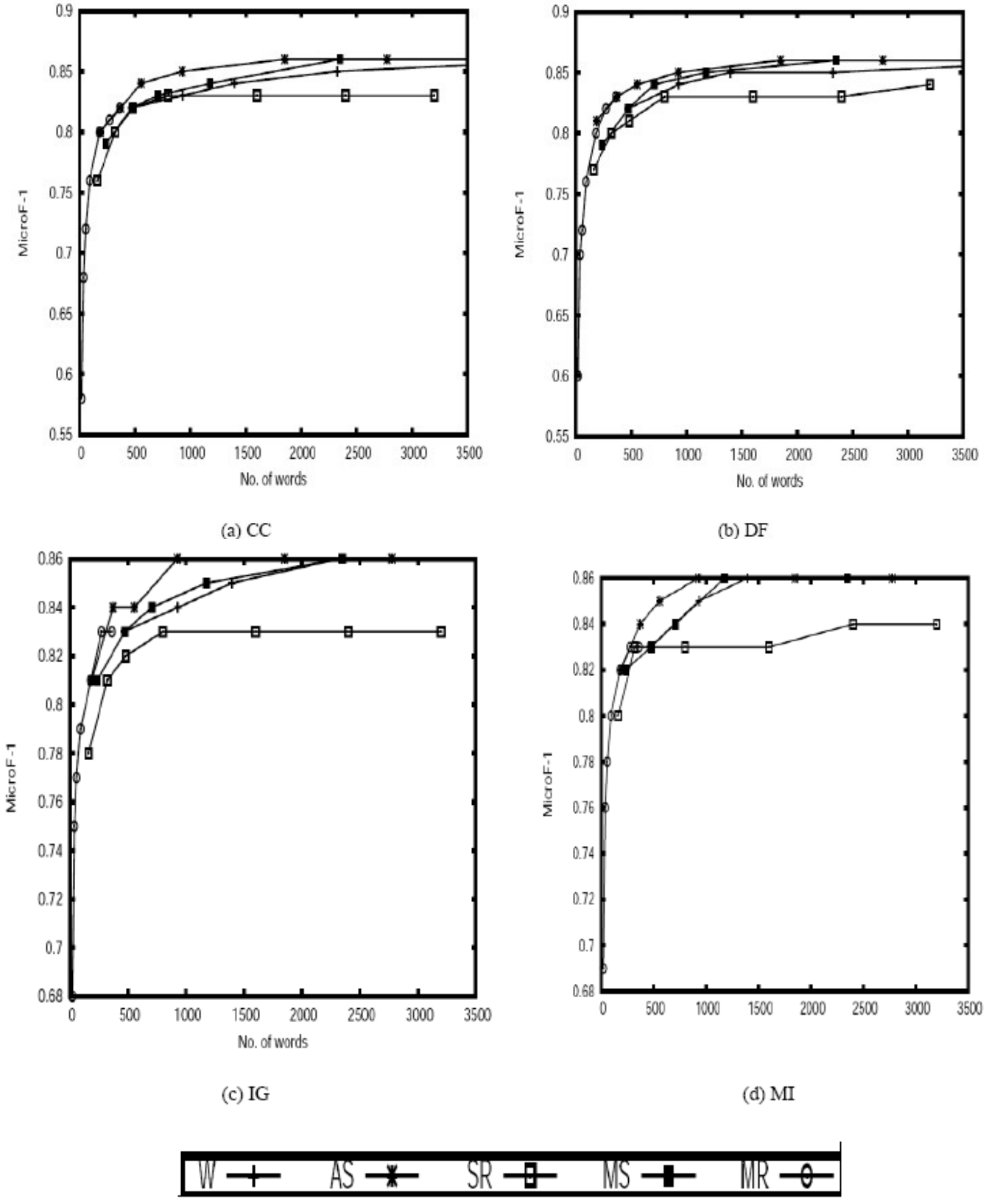


Figure 3 Average $MicroF_1$ of five random splits of Alj-Mgz dataset using (a) CC; (b) DF; (c) IG; and (d) MI feature scoring methods with different threshold values (0.5% – 10%) of unique words in the dataset where W is using raw word, AS is Al-Stem stemmer, SR is Sebawai root extractor, MS is RDI MORPHO3 stemmer, and MR is RDI MORPHO3 root extractor.

4. Results

Figure 1 illustrates the size of Alj-News and Alj-Mgz datasets according to the preprocessing tool used. Using a root extraction tool leads to a significant decrease in the vocabulary size compared with using a stemmer. On the other hand, using the raw text leads to the largest vocabulary size especially for large threshold values.

Additionally, the AS stemmer reduces the vocabulary size much more compared with the MS stemmer, while the MR root extractor was better than SR in terms of the vocabulary size. In the following, we will focus on the evaluation of the performance of the preprocessing tools for each dataset separately.

Alj-News Dataset

Figure 2 represents theMicroF1 of Alj-News dataset. The results show that the worst performance is obtained when using the raw text. The performance of the four other preprocessing tools is nearly identical when using CC and DF feature scoring methods. However, the superiority of AS is apparent when using IG and MI scoring methods. This is mainly because the performance of the IG and MI method is better than the CC and DF in all threshold values. This may indicate that using light stemmers such as AS with a good feature selection method may boost the performance of TC even when using small number of features (e.g. 5%).

Alj-Mgz Dataset

Figure 3 shows the average performance of five splits of Alj-Mgz dataset using MicroF1. The Alj-Mgz dataset differs from the Alj-News dataset in three characteristics. First, the size of the Alj-Mgz is much larger than the Alj-News dataset. Second, the distribution of documents among categorizes in Alj-Mgz is skewed while the documents of Alj-News dataset is evenly distributed. Finally, the classification task of the Alj-Mgz dataset is much harder than the Alj-News dataset. Due to all these factors, the conclusion derived from the Alj-Mgz is different than those obtained from the Alj-News dataset. The results show that the SR tool leads to the worst performance in all feature selection methods. On the other hand, the performance of the MR and AS tools was nearly superior. As the threshold increases, the performance of the two stemmers and the raw text becomes identical for IG and MI feature scoring methods. It should be noted that the performance due to using the raw text is slightly worse than the performance obtained when using the two stemmers for CC and DF scoring methods.

5. Conclusions

This work conducts an evaluation study among several preprocessing tools in Arabic TC. We compared the raw text (W), Al-Stem stemmer (AS), Sebawai root extractor (SR), RDI MORPHO3 stemmer (MS), and RDI MORPHO3 root extractor (MR). The study was performed using four feature scoring methods and different threshold values. Two datasets were used in this study; namely Alj-News, and Alj-Mgz datasets. These datasets are of different nature in terms of the difficulty of the classification task, the document distribution among categories, and vocabulary size. The results show that using light stemmer (AS) with a good-performing feature selection method such as MI or IG enhances the performance for small sized datasets and small threshold values for large datasets. Additionally, using the raw text leads to the worst performance in small datasets while its performance was among the best tools in large datasets. This may explain the contradiction in the results obtained previously in the literature of the Arabic text categorization since the performance of the preprocessing tools is affected by the characterizes of the dataset used.

Another contribution of this work is investigating the effect of the preprocessing tools using different feature selection methods. To the best of our knowledge, such study has not been performed on the Arabic language.

The results show that the stemmer may enhance the performance, compared with raw text, even if the feature scoring method used is poorly performing.

A final contribution of this paper is the comparison it provides among two well-known Arabic morphological tools. The results showed that the light stemmer Al-Stem performed better than MORPHO3 stemmer while MORPHO3 root extractor is better than Sebawai root extractor. Additionally, the results showed that Al-Stem leads to a reduced vocabulary size compared with MORPHO3 stemmer while MORPHO3 root extractor provides a less vocabulary size compared with Sebawai.

Acknowledgements

We would like to thank MEDAR for providing a travel grant for the first author of this paper. Additionally, we thank RDI for providing us with their morphological analyzer (MORPHO3). Thanks should go also to Dr. Kareem Darwish for making Al-Stem and Sebawai available for research use. Additionally, we would like to thank the authors of (Mohamed et al., 2005) for providing us with Alj-News dataset. We also appreciate the efforts made by Mr. Kareem Said, Mr. Mahmod Said, and Miss Nora Bilal in collecting Alj-Mgz dataset.

Bibliographical References

Attia, M. (2000). A large-scale computational processor of the Arabic morphology. Master's thesis, Computer Engineering, Faculty of Engineering, Cairo, Egypt.

Bekkerman, R., El-Yaniv, R., Tishby, N., & Winter, Y. (2003). Distributional word clusters vs. words for

text categorization. *Journal of Machine Learning Research*, 3, 1183–1208.

Brants, T., Chen, F., & Farahat, A. (2002). Arabic document topic analysis. In *Proc. of the LREC-2002 Workshop Arabic Language Resources and Evaluation* Las Palmas, Spain.

Darwish, K. (2002). Building a shallow Arabic morphological analyzer in one day. In *Proc. Of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)* (pp. 1–8). Philadelphia, Pennsylvania, United States.

Darwish, K. (2003). *Probabilistic methods for searching OCR-degraded Arabic text*. PhD thesis, University of Maryland, College Park, Maryland, United States.

Darwish, K., Hassan, H., & Emam, O. (2005). Examining the effect of improved context sensitive morphology on Arabic information retrieval. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 25–30). Ann Arbor,Michigan.

Debole, F. & Sebastiani, F. (2005). An analysis of the relative hardness of reuters-21578 subsets. *Journal of the American Society for Information Science and Technology (JASIST)*, 56(6), 584–596.

Joachims, T. (1999). Making large-scale support vector machine learning practical. In *Advances in Kernel Methods – Support Vector Learning* (pp. 169–184).

Larkey, L., Ballesteros, L., & Connell, M. (2002). Improving stemming for Arabic information retrieval: light stemming and occurrence analysis. In *Proc. of the 25th ACM International Conference on Research and Development in Information Retrieval (SIGIR'02)* (pp.275–282). Tampere, Finland.

Lewis, D. & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proc. Of the 3rd Symposium on Document Analysis and Information Retrieval (SDAIR'94)* (pp. 81–93). Las Vegas, United States: ISRI; University of Nevada.

Mohamed, S., Ata,W., & Darwish, N. (2005). A new technique for automatic text categorization for Arabic documents. In *Proc. of the 5th IBIMA International Conference on Internet and Information Technology in Modern Organizations* Cairo, Egypt.

Moukdad, H. (2006). Stemming and root-based approaches to the retrieval of Arabic documents on the web. *Webology*, 3(1). Article 22.

Ng, H., Goh,W., & Low, K. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In *Proc. of the 20th ACM International Conference on Research and Development in Information Retrieval (SIGIR'97)* (pp. 67–73). Philadelphia, United States.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1), 1–47.

Silvatt, C. & Ribeirot, B. (2003). The importance of stop word removal on recall values in text categorization. In *Proc. of the IEEE International Joint Conference on Neural Networks (IJCNN 2003)*, volume 3 (pp. 1661–1666). Portland, Oregon, USA.

Song, F., Liu, S., & Yang, J. (2005). A comparative study on text representation schemes in text categorization. *Pattern Analysis & Applications*, 8(1-2), 199–209.

Yang, Y. & Pedersen, J. (1997). A comparative study on feature selection in text categorization. In *Proc. Of the 14th International Conference on Machine Learning (ICML'97)* (pp. 412–420). Nashville, Tennessee, United States.

Zhu, M., Zhu, J., & Chen,W. (2005). Effect analysis of dimension reduction on support vector machines. In *Proc. of the Natural Language Processing and Knowledge Engineering IEEE NLP-KE* (pp. 592–596).Wuhan, China.