International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.67-80
http://dx.doi.org/10.14257/ijdta.2016.9.6.07

The Impact of Feature Reduction Techniques on Arabic Document Classification

Abdullah Ayedh[1], Guanzheng Tan[2*] and Hamdi Rajeh[3]

[1,2]*School of Information Science and Engineering, Central south University, Changsha, Hunan (HN), 410000/ Time (UTC+8), China*
[3]*School of Information Science and Engineering, Hunan University, China*
[1]*abdullah_ayedh@csu.edu.cn;* [2]*tgz@csu.edu.cn;* [3]*hamdiahmed919@gmail.com*

Abstract

Feature reduction are common techniques that used to improve the efficiency and accuracy of the document classification systems. The problems associated with these techniques are the highly dimensionality of the feature space and The difficulty of selecting the important features for understanding the document in question. The document usually consists of several parts and the important features that more closely associated with the topic of the document are appearing in the first parts or repeated in several parts of the document. Therefore, the position of the first appearance of a word and the compactness of the word considered as factors that determine the important features using the information within a document. This study, explored the impact of combining three feature weighting methods that depend on inverse document frequency (IDF), namely, Term frequency (TFiDF), the position of the first appearance of a word (FAiDF), and the compactness of the word (CPiDF) on the classification accuracy. In addition, we have investigated different feature selection techniques, namely, Information gain (IG), Goh and Low (NGL) coefficients, Chi-square Testing (CHI), and Galavotti-Sebastiani-Simi Coefficient (GSS) in order to improve the performance for Arabic document classification system. Experimental analysis on Arabic datasets reveals that the proposed methods have a significant impact on the classification accuracy, and in most cases the FAiDF feature weighting performed better than CPiDF and TFiDF. The results also clearly showed the superiority of the GSS over the other feature selection techniques and achieved 98.39% micro-F1 value when using a combination of TFiDF, FAiDF, and CPiDF as feature weighting method.

Keywords: *Feature selection; Feature representation; Feature weighting; Document categorization, dimensionality reduction*

1. Introduction

Recently, the web is the main source of unstructured and semi structured information include the governmental repositories, news articles, biological databases, chat rooms, digital libraries, online forums, E-mail and blog repositories. Researchers are highly challenged to find better ways to deal with such huge amount of information in order to provide relevant information accurately. The task of automatic document classification became one of the key methods for organizing the information and knowledge discovery. Natural language processing (NLP), data mining, and machine learning techniques are used to implementation of any document classification systems. Therefore, there are several challenges, such as appropriate document representation, dimensionality reduction to handle algorithmic issues and an appropriate classifier to obtain good generalization and avoid over-fitting [1].

* Corresponding Author

ISSN: 2005-4270 IJDTA

The document in text classification system usually pass through three main stages document preprocessing, document modeling, and document classification. The first stage

involves some feature extraction techniques such as tokenization, stop word removal, normalization, and stemming. The second stage includes feature selection, feature representation, and feature weighting. The final stage is document classification wherein documents are divided into training and testing data. In this stage, the proposed classification algorithm uses the training data to obtain a classification model that will be evaluated by means of the testing data.

The high dimensionality of the feature space is a major challenge for many classification techniques and increases the complexity of many classification algorithms. Methods and techniques that can improve the performance and efficiency of the classification system by reducing the data into small dimensional space are highly desired [2]. Feature selection and feature weighting are common techniques and methods that are used in document classification to reduce the highly dimensionality of the feature space and to improve the efficiency and accuracy of the classification system.

The main contribution of this paper is using new feature weighting methods that depend on inverse document frequency (IDF). The proposed methods take into account the important of the first appearance of a word and the compactness of the word which can be taken as factors that determine the important features in the document. The motivation behind this consideration by the fact that, the document is written in an organized manner to describe its main topic(s). It's usually consisting of several parts and the important features that more closely associated with the topic of the document are appearing in the first parts or repeated in several parts of the document. For example, the main topic for news articles may mentions at the title and the first part of the document to draw the attention of the reader. Therefore, depending on the location, the document parts may have different degrees of contribution to the document's main topic(s) [3]. In addition, we have investigated different feature selection methods that have a significant impact on reducing the dimensionality of feature space and thus improve the performance of Arabic document classification system.

The rest of this paper is organized as follows. In Section 2, Related works are presented. The methodology adopted in this study is elaborated in Section 3. Experiments and results are presented in Section 4. Finally, conclusions and future works are provided in Section 5.

2. Related Works

Several studies have been conducted to evaluate the impact of feature reduction techniques on document classification. This section summarizes what has been achieved on document classification from various pieces of the literature.

Mesleh[4] studied the impact of six commonly used feature selection techniques namely, CHI, NGL, GSS, IG, odd ratio (OR) and mutual information (MI) on Arabic document classification. His experiments are performed using support vector machine (SVM) and showed that CHI, NGL and GSS significantly outperforms the other techniques and achieved high classification efficiency in terms of the F-measure (88.11%).

Syiam *et al.* [5] examined several feature selection approaches on Arabic document classification. the authors recommended to address the problem of Arabic document categorization using n-gram statistical indexing for document preprocessing, hybrid approach of document frequency thresholding (DF), information gain (IG) for feature selection, normalized-TFiDF for feature weighting and Rocchio classifier for classification. Experimental results demonstrate the effectiveness of the proposed model and gives generalization accuracy of about 98%.

Khorsheed and Al-Thubaity [6] investigated classification techniques with a large and diverse dataset. These techniques include a wide range of classification algorithms, feature selection methods, and representation schemes. For feature selection, their best average result was achieved using the GSS method with TF as the base for calculations. For feature weighting functions, the study concludes that length term collection (LTC) was the best performer, followed by Boolean and term frequency collection (TFC). His experiments showed that the SVM classifier outperformed the other algorithms.

Chirawichitchai *et al.* [7] compared the efficiency of several feature representation schemes, including Boolean, TF, TFiDF, TFC, LTC, entropy, and Term Frequency Relevance Frequency (TF-RF) on Thai Document Categorization. After running the experiments on Thai news article corpus with three supervised learning classifiers, including SVM, naïve bayes (NB) and decision tree (DT) classifiers, the authors claimed that using TF-RF weighting with SVM classifier yielded the best performance with the F-measure equaling 95.9%.

Saad EM *et al.* [8] presented new semantic feature reduction approach which is based on synonyms merge to preserve features semantic and prevent important terms from being excluded. In this approach, five feature selection methods were applied after synonym merges, DF, TFiDF, CHI, IG, and MI to produce a more compact feature space. Experiments shows that classification performance is increased after merging terms and yielding best performance for CHI and IG selection methods.

Leena. H. Patil *et al.* [9] proposed multistage feature selection model for document classification using IG and Rough set. Experiments are performed using k-nearest neighbour (KNN) and NB classifier on Reuters 21578, Classic 04 and News Group 20. The study concluded that a multistage feature selection model for document classification using IG and Rough set is efficient to reduce the dimensionality of feature space.

Ababneh *et al.* [10] discussed different variations of vector space model (VSM) to classify Arabic documents using KNN algorithm, these variations are Cosine coefficient, Dice coefficient and Jacaard coefficient and using IDF term weighting method for comparison purposes. The experimental results showed that Cosine coefficient outperforms Dice and Jaccard coefficients.

Zahran and Kanaan [11] introduced feature selection algorithm based on particle swarm optimization (PSO) to improve the efficiency of Arabic document classification. The experimental results showed the superiority of the proposed algorithm compared with DF, TFiDF and CHI in terms of classification accuracy.

Zaki *et al.* [12] proposed a hybrid system for Arabic document classification based on the semantic vicinity of terms and the use of a radial basis modeling. They use the hybridization of n-grams and the TFiDF measure to calculate the similarity between words. By comparing the obtained results, they found that the use of radial basis functions improve the performance of the Arabic document classification system.

Aymen Abu-Errub [13] introduced a new method Arabic text classification algorithm using TFiDF, and CHI methods. The researcher examined the proposed algorithm using 1090 documents categorized into ten main categories and 50 sub categories. The experimental results showed that the proposed algorithm is capable of classifying the tested documents to its appropriate sub category.

The conclusion after looking at the related works in this area, there is no superior feature reduction method for all datasets, the feature reduction methods which are reported as the best for English document classification are not the best for Arabic document classification, and the results are varied and are not consistent for all studies.

As a result, the proposed methods which depend on the compactness of the appearances of a word and the position of the first appearance of a word, could be considered as a new feature weighting methods for Arabic document classification using the information within a document.

3. Methodology

An Arabic document classification system usually consists of three main stages: preprocessing stage, document modeling stage, and document classification stage. The preprocessing stage involves some feature extraction techniques such as tokenization, stop word removal, normalization, and stemming. Document modeling stage which also known as "dimensionality reduction" includes feature selection, feature representation, and feature weighting. Document classification stage covers classification model construction and classification model evaluation (Figure 1). These phases will be described in details in the following subsections.

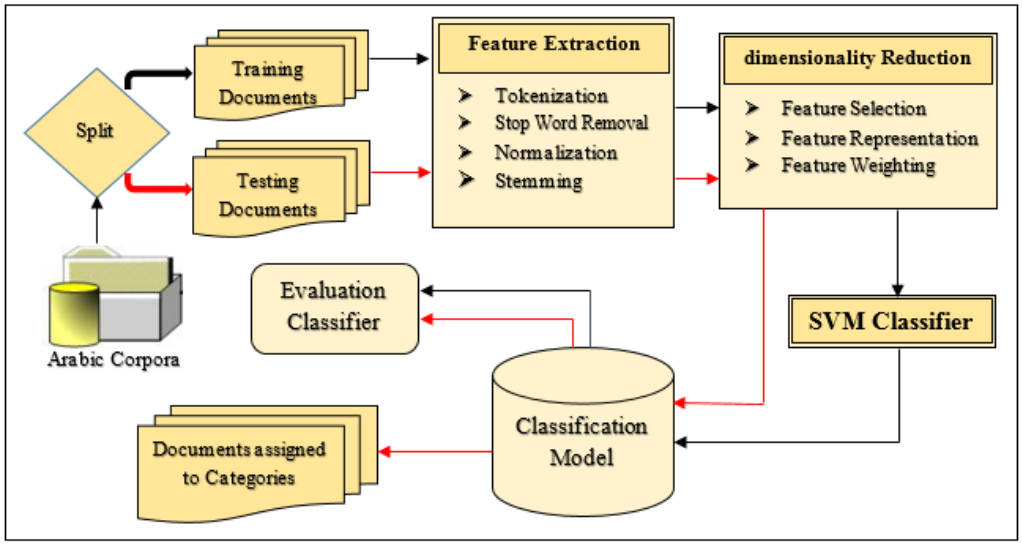

Figure 1. Arabic Document Classification System

3.1. Feature Extraction

Feature extraction is one of the most important stages in document classification systems especially with highly inflected languages like Arabic language. Feature extraction techniques used to convert the document from original data source into a format which suitable for classification purpose [14].

The most popular feature extraction techniques for Arabic document classification are tokenization, stop word removal, normalization, and stemming. Tokenization is a method for dividing texts into tokens. These tokens could be individual words that are converted without understanding their meanings or relationships. Stop word removal used to purge the features from noise such as articles, prepositions, conjunctions, pronouns, numbers, non-Arabic letters, and symbols. Normalization aims to normalize certain letters that have different forms in the same word to one form. Stemming method is used to reduce different forms of the feature that reflect the same meaning to single form (its root or stem).

3.2. Feature Selection

A large number of features that extracted from feature extraction stage are irrelevant to the classification task and can be removed without affecting the classification accuracy for several reasons: First, the performance of some classification algorithms is negatively affected when dealing with a high dimensionality of features. Second, an over-fitting problem may occur when the classification algorithm is trained in all features. Finally,

some features are common and occur in all or most of the categories. The mechanism that removes the irrelevant feature is called feature selection [6].

Feature selection can be defined as the process of selecting the most representative subset that contains the most relevant features for each category in the training set based on a few criteria and using this subset in document classification [15]. Feature selection can be local or global. Global feature selection consists of generating a subset of features from all categories, while local feature selection creates a subset for each document category, where the most relevant features of the category are included[16].

There are many feature selection methods have been introduced for document classification. The most frequently used methods have been Document Frequency Threshold (DF), Information Gain (IG), Chi-square Testing (x^2), mutual information (MI), odds ratio (OR), Goh and Low (NGL) coefficients, Darmstadt indexing approach (DIA), and Galavotti-Sebastiani-Simi Coefficient (GSS) [17]. All these methods order the features according to their importance or relevance to the category. The top rank futures from each category are then chosen and represented to the classification algorithm.

In this study, we investigated four feature selection methods for Arabic documents classification, namely, Information gain (IG), Goh and Low (NGL) coefficients, Chi-square Testing (x^2), and Galavotti-Sebastiani-Simi Coefficient (GSS). The next section explains these methods in more details.

Information Gain (IG)

IG is commonly used as a feature goodness criterion in machine learning. IG measures the amount of information obtained for category prediction by knowing the presence or absence of a feature in a document [18-19]. the IG idea is to determine features that reveal the most information about the categories. The IG of feature *t* is defined as:

$$\mathrm{IG}(\mathrm{t},\mathrm{c}_i) = \sum_{i=1}^{i=m} P(t,c_i)\cdot\log\frac{P(t,c_i)}{P(t)\cdot P(c_i)} + \sum_{i=1}^{i=m} P(\bar{t},c_i)\cdot\log\frac{P(\bar{t},c_i)}{P(\bar{t})\cdot P(c_i)} \quad (1)$$

And is estimated using:

$$\mathrm{IG}(\mathrm{t}) = \sum_{i=1}^{i=m} A\cdot\log\frac{A}{(A+C)(A+B)} + \sum_{i=1}^{i=m} B\cdot\log\frac{B}{(B+D)(A+B)} \quad (2)$$

Chi-square Testing (x^2)

Chi-square testing (x^2) is a well-known discrete data hypothesis testing method from statistics, which evaluates the correlation between two variables and determines whether they are independent or correlated [20]. The chi-square statistics show us the relevance of each feature to the category. The value of x^2 for each feature *t* in a category *c* can be defined by equation (3,4) [17].

$$\mathrm{x}^2(\mathrm{t}_k,\mathrm{c}_i) = \frac{\left|Tr\right|\cdot\left[P(t_k,c_i)\cdot P(\bar{t_k},\bar{c_i}) - P(t_k,\bar{c_i})\cdot P(\bar{t_k},c_i)\right]^2}{P(t_k)\cdot P(\bar{t_k})\cdot P(c_i)\cdot P(\bar{c_i})} \quad (3)$$

and is estimated using:

$$\mathrm{x}^2(\mathrm{t},\mathrm{c}) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (4)$$

Ng-Goh-Low (NGL) Coefficient

NGL Coefficient is a variant of x^2 metric. A positive NGL value indicates that word is a possible feature and correlates with category c while a negative value means word correlates with category $\bar{c}$ [21]. The NGL value can be computed as follows:

$$NGL(t_k, c_i) = \frac{\sqrt{Tr} \cdot [P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)]}{\sqrt{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}} \quad (5)$$

Which we can easily compute using the *A, B, C, D* values as:

$$NGL(t, c) = \frac{\sqrt{N} \cdot (AD - CB)}{\sqrt{(A + C)(B + D)(A + B)(C + D)}} \quad (6)$$

Galavotti-Sebastiani-Simi (GSS) Coefficient

GSS propose a simplified x^2 statistic. They remove the $\sqrt{n}$ factor, and the denominator completely. They describe the $\sqrt{n}$ factor as being unnecessary. They also remove the denominator, $\sqrt{(A + C)(B + D)(A + B)(C + D)}$, by giving the reason that the denominator gives high Correlation Coefficient score to rare words, and rare categories [21]. The GSS value can be computed as follows:

$$GSS(t, c_i) = P(t, c_i) \times P(\bar{t}, \bar{c}_i) - P(t, \bar{c}_i) \times P(\bar{t}, c_i) \approx \frac{AD - CB}{N^2} \quad (7)$$

For all above feature selection methods, *A* is the number of documents of category *c* containing the feature *t*; *B* is the number of documents of other category (not *c*) containing *t*; *C* is the number of documents of category *c* not containing the feature *t*; *D* is the number of documents of other category not containing *t*; *N* is the total number of documents.

3.3. Feature Representation

The selected features from the previous step are needed to be represented in a form that is suitable for classification tasks. A well representation for features leads to gain more accuracy for document classification system.

In this study, we divided a document into several passages; Each passage is represented as a vector, where each of its dimension corresponds to a separate word in the document collection. If a word occurs in the document, the number of appearances of this word in the corresponding part is non-zero.

Kim and Kim [3] discussed three type of passages to split document, namely, discourse passages, semantic passages, and window passages. In this study, we adopted non-overlapping window passages which divide document into a fixed-length sequence of words without sharing any words at the boundary between two adjacent passages.

Several different methods used for computing weight for each feature in the VSM. the best known is TFiDF where TF is frequency of term *t* in document *d*, and inverse document frequency (IDF) measures the global relevance of the word within a collection of documents.

In addition to term frequency TF in TFiDF method the compactness of appearance of the word and the position of the first appearance of the word can be considered as a criterion that measure the importance of a feature in a document [22]. Therefore, the standard TFiDF formula can be generalized as follows:

$$\mathrm{TFiDF}(t,d) = Importance(t,d) \times idf(t) \tag{8}$$

The *importance (t, d)* corresponds to different functions, the term frequency (TF), the compactness of the appearances of a word (CP), and the position of the first appearance of a word FA. TF, CP, and FA are calculated as follows:

Term Frequency Related to its Inverse Document Frequency (TFiDF)

The first criterion is the term frequency (TF) in the given document which offers a measure of the relevance of the feature within a document. The TFiDF method uses the TF and document frequency (DF) to compute the weight of a word in a document by these formulas (9,10,11) [23]:

$$\mathrm{TF}(t,d) = \frac{count(t,d)}{size(d)}, \tag{9}$$

$$\mathrm{IDF}(t) = \log\left(\frac{N}{DF(d,t)}\right) \tag{10}$$

$$\mathrm{TF/IDF}(t,d) = TF(t,d) * IDF(t) \tag{11}$$

In the above formulas (9,10), *count (t, d)* is the number of times term *t* occurs in document *d*, *Size(d)* is the total number of words of document *d*, DF (d, t) is the number of documents that contain term *t*, and *N* is the total of all documents in the training set.

Compactness of Appearance of the Word

The second criterion is the compactness which measures whether the appearances of a word concentrated in a certain part of a document or distributed over the whole document. In the first case, the word described as more compact and in the second case the word described as less compact. The motivation behind this consideration that whenever the word is more likely to appear in several parts this indicates that this word is more closely associated with the topic of the document.

To measure the compactness of the appearances of a word, the variance of the positions of all appearances CP_{PosVar} is used. The TFiDF depending on the compactness of the appearances of the word (CP) weighting method is defined by these formulas (12,13,14,15):

$$\mathrm{count}(t,d) = \sum_{i=0}^{n-1} c_i, \tag{12}$$

$$\mathrm{CP}_{\mathrm{PosVar}}(t,d) = \frac{\sum_{i=0}^{n-1} c_i \times \left| i - \frac{\sum_{i=0}^{n-1} c_i \times i}{count(t,d)} \right|}{count(t,d)} \tag{13}$$

$$\mathrm{CP}(t,d) = \frac{CP_{PosVar}(t,d) + 1}{len(d)}, \tag{14}$$

$$\mathrm{TFiDF}(t,d) = CP(t,d) * IDF(t) \tag{15}$$

Where *len(d)* is the total number of parts of document *d*.

Position of First Appearance of the Word

The third criterion is the position of the first appearance of a word. The motivation behind this consideration is by the fact that, the important contents that more closely associated with the topic of the document are appearing in the first parts of a document.

The position of the first appearance $FirstApr$ can be extracted directly from the proposed VSM by this formula (16):

$$\mathrm{FirstApr}(t,d) = \min \min_{i \in \{0 \ldots n-1\}} c_i > 0 \,?\, i : n, \tag{16}$$

Then, the TFiDF depending on the first appearance of the word (FA) weighting method is defined by this formulas (17,18):

$$\mathrm{FA}(t,d) = \frac{\left| FirstApr(t,d) - \frac{len(d)-1}{2} \right| + 1}{len(d)} \tag{17}$$

$$\mathrm{TFiDF}(t,d) = FA(t,d) * IDF(t) \tag{18}$$

3.4. Classification Algorithm

The documents can be classified by three ways, unsupervised, supervised and semi supervised methods. Many techniques and algorithms are proposed recently for automatic document classification. Normally supervised learning techniques are used for automatic document classification, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents [1] such as naïve bayes (NB), k-nearest neighbor (KNN), and support vector machine (SVM) *etc.*

SVM has become a popular algorithm in the last years due to its good performance. Furthermore, it can handle documents with high-dimensional input space, and culls out most of the irrelevant features [24]. So in this study, SVM classifier was applied to observe the Impact of feature reduction techniques on improving classification accuracy.

4. Experiment and Results

4.1. Data Collection

To evaluate the impact of feature reduction techniques on the accuracy of Arabic document classification, we have used an in-house dataset collected from several published papers for Arabic document classification and from scanning the well-known and reputable Arabic websites. The collected corpus contains 32090 documents divided into nine categories of News, Economy, Health, History, Sport, Religion, Social, Nutriment, and Law that vary in length and number of documents. The statistics of the corpus are shown in Table 1.

Table 1. Statistics of the Documents in the Corpus

Category Name	Number of Documents
News	6860
Economy	4780
Health	2590
History	3230
Sport	3950
Religion	3470
Social	3600
Nutriment	2370
Law	1240
Total	**32090**

4.2. Experimental Configuration and Performance Measure

In this study, the benchmarking dataset mentioned in the previous section need a set of preprocessing routines to be suitable for classification purpose. All documents in the dataset were prepared by converting them to UTF 8 encoding. For preprocessing process, tokenization, stop words removal, normalization, and stemming were used. For stop word removal a list of 896 words was prepared to be eliminated from all the documents. In this study, the linear kernel for SVM classifier was applied because it has been clarified that the most classification problems are linearly separable[24].

Cross-validation was used for all classification experiments, which partitioned the complete collection of documents into 10 mutually exclusive subsets called folds. Each fold has the same number of documents. One of the subsets is used as the test set, whereas the rest of the subsets are used as training sets.

The evaluation of the performance for classification model to classify documents into the correct category is conducted by using several mathematical rules such as recall (R), precision (P), and F-measure (F), which are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

Where *TP* is the number of documents that are correctly assigned to the category, *TN* is the number of documents that are correctly assigned to the negative category, *FP* is the number of documents that are incorrectly assigned to the category by the system, and *FN* is the number of documents that belong to the category but are not assigned to the category.

The success measure, namely, micro-F1 score, a well-known F1 measure, is selected for this study, which is calculated as follows:

$$\text{Micro-F1} = \frac{2 * Precision \times Recall}{Precision + Recall} = \frac{2 \cdot \sum_{k=1}^{c} TP_k}{\sum_{k=1}^{c} FP_k + 2 \cdot \sum_{k=1}^{c} TP_k + \sum_{k=1}^{c} FN_k} \tag{21}$$

4.3. Results and Analysis

In the following experiment we studied the impact of feature reduction on Arabic document classification accuracy. Three feature representation schemas, namely, TFiDF, FAiDF, and CPiDF and four feature selection methods, namely, IG, chi-square (CHI), NGL, and GSS are used in this experiment. The classification accuracy for SVM classifier based on three feature representation schemas, four feature selection methods and three n ranked thresholds, are shown in Table 2 and Figure 2.

Table 2. Micro-F1 Scores for SVM Classifier Based on Four Feature Selection, and Three Feature Representation Schemas

SVM	CHI			IG			NGL			GSS		
	100	500	1000	100	500	1000	100	500	1000	100	500	1000
TFiDF	0.8715	0.9255	0.9482	0.8008	0.9195	0.9315	0.7431	0.9130	0.9253	0.8423	0.9298	0.9520
FAiDF	0.8422	0.9548	0.9636	0.8046	0.9296	0.9479	0.7609	0.9170	0.9353	0.8676	0.9582	0.9709
CPiDF	0.8784	0.9358	0.9537	0.818	0.9279	0.9416	0.7572	0.9077	0.9324	0.8993	0.9403	0.9611

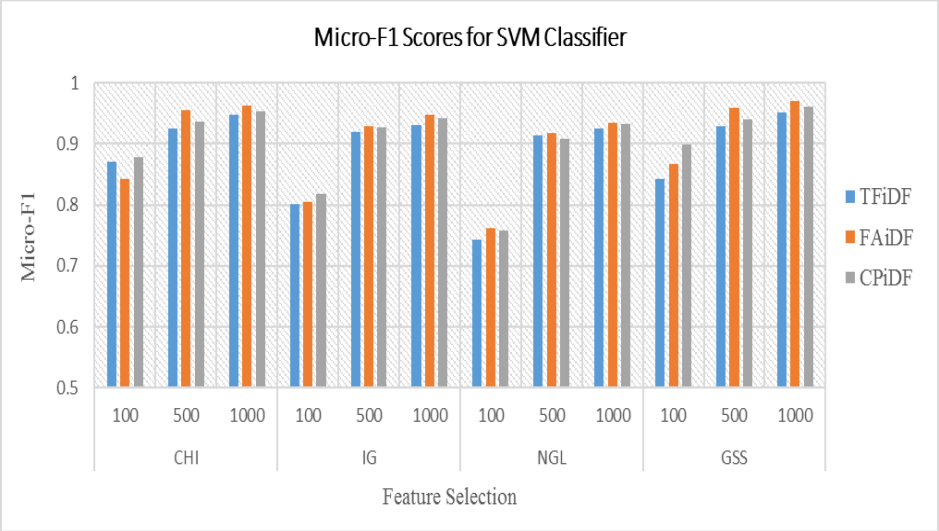

Figure 2. Micro-F1 Scores for SVM Classifier Based on Four Feature Selection, and Three Feature Representation Schemas

The maximum accuracy achieved is 97.09% micro-F1 using FAiDF and GSS when features size was 1000. The results showed that the CPiDF and FAiDF both significantly improved the baseline, and in most cases the FAiDF feature representation performed better than CPiDF and TFiDF. On the other hand, the GSS and Chi-square feature selection methods achieved comparable accuracy and the highest accuracy is achieved when one of them is used.

As a result, choosing the appropriate features selection methods and feature representation schemas are indeed helpful to improve the efficiency of document classification system and provides significant improvement on the accuracy of Arabic document classification.

Combining Feature Weighting Methods

We also studied the impact of combining feature weighting methods on the accuracy of Arabic document classification. All possible combinations of the feature weighting methods listed in Table 3 are considered during the experiments to reveal all possible interactions between the feature weighting methods.

First appearance (FR the table) is either 0 or 1; that is, first appearance method is ignored or used. Compactness (CP in the table) is either 1 or 0; that is, compactness method is used or not. Term frequency (TF in the table) is either 1 or 0; that is, term frequency method is used or not. We used GSS feature selection method, and SVM classifier in this experiment.

Table 3. Combinations of Feature Weighting Methods

No.	Feature Weighting Methods Combinations
1	First Appearance (FA):1 | Compactness (CP):1 | Term Frequency (TF):0
2	First Appearance (FA):1 | Compactness (CP):0 | Term Frequency (TF):1
3	First Appearance (FA):0 | Compactness (CP):1 | Term Frequency (TF):1
4	First Appearance (FA):1 | Compactness (CP):1 | Term Frequency (TF):1

The results of the experiments of all possible combinations of the three feature weighting methods using SVM algorithm are illustrated in Table 4 and Figure 3.

Table 4. Micro-F1 Scores for SVM Classifier based on GSS Feature Selection, and Three Feature Weighting Methods Combination

Classifier	GSS Feature Selection			Feature Weighting Combination		
	100	500	1000	TFiDF	FAiDF	CPiDF
SVM	0.8423	0.9298	0.9520	1	0	0
	0.8676	0.9582	0.9709	0	1	0
	0.8993	0.9403	0.9611	0	0	1
	0.8737	0.9675	0.9768	1	1	0
	0.9017	0.9443	0.9675	1	0	1
	0.9059	0.9620	0.9783	0	1	1
	0.9072	0.9692	0.9839	1	1	1

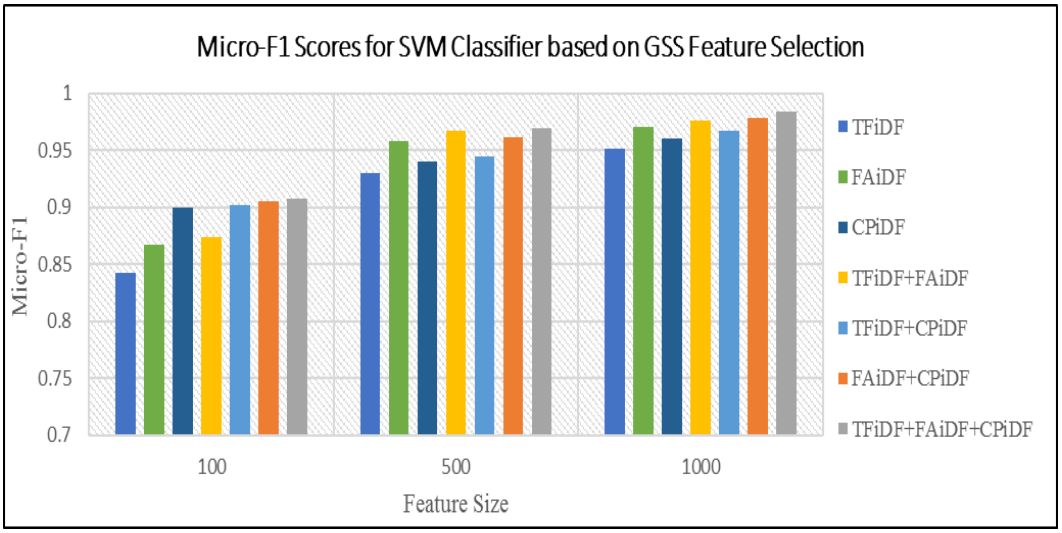

Figure 3. Micro-F1 Scores for SVM Classifier Based on GSS Feature Selection, and Three Feature Weighting Methods Combination

The results showed that combining two or three feature weighting methods showed significant improvement in classification accuracy. The maximum improvement was 1.3% when TFiDF, FAiDF, and CPiDF were combined using SVM classifier, and GSS feature selection method when feature size was 1000.

5. Conclusion

In this study, we investigated the impact of feature reduction methods on the accuracy of Arabic document classification. The document usually consists of several parts and the important features that more closely associated with the topic of the document are appearing in the first parts or repeated in several parts of the document. Therefore, the position of the first appearance of a word and the compactness of the word considered as a factors that determine the important features using the information within a document.

Our methodology to represent a document in feature space depends on dividing a document into several passages; Each passage is represented as a vector, where each of its dimension corresponds to a separate word in the document collection. We used a combination of three feature weighting methods, namely, term frequency (TFiDF), the position of the first appearance of a word (FAiDF), and the compactness of the word (CPiDF). In addition, we have investigated four feature selection techniques, namely, Information Gain (IG), Goh and Low (NGL) coefficients, Chi-square Testing (x^2), and Galavotti-Sebastiani-Simi Coefficient (GSS) in order to improve the performance for Arabic document classification system.

The results obtained from the experiments reveal that CPiDF and FAiDF both significantly improved the baseline, and in most cases the FAiDF feature representation schemas performed better than CPiDF and TFiDF. Combining two or three feature weighting methods showed insignificant improvement in classification accuracy. The maximum improvement was 1.3% when TFiDF, FAiDF, and CPiDF feature weighting methods were combined using SVM classifier, and GSS feature selection.

The results also clearly showed the superiority of the GSS over the other feature selection techniques. GSS feature selection technique achieved 98.39% micro-F1 value when using a combination of TFiDF, FAiDF, and CPiDF as feature weighting method.

References

[1] B. Baharudin, L. H. Lee and K. Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal of Advances in Information Technology, vol. 1, no. 1, **(2010)**.
[2] J. Yan, "OCFS: optimal orthogonal centroid feature selection for text categorization", in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, **(2005)**.
[3] J. Kim, and M. H. Kim, "An evaluation of passage-based text categorization", Journal of Intelligent Information Systems, vol. 23, no. 1, **(2004)**, pp. 47-65.
[4] A. Moh'd Mesleh, "Support vector machines based Arabic language text classification system: feature selection comparative study", in Advances in Computer and Information Sciences and Engineering, Springer, **(2008)**, pp. 11-16.
[5] M. M. Syiam, Z. T. Fayed and M. B. Habib, "An intelligent system for Arabic text categorization", International Journal of Intelligent Computing and Information Sciences, vol. 6, no. 1, **(2006)**, pp. 1-19.
[6] M. S. Khorsheed, and A. O. Al-Thubaity, "Comparative evaluation of text classification techniques using a large diverse Arabic datase", Language resources and evaluation, vol. 47, no. 2, **(2013)**, pp. 513-538.
[7] N. Chirawichitchai, P. Sa-nguansat and P. Meesad, "Developing an effective Thai Document Categorization Framework base on term relevance frequency weighting", in Knowledge Engineering, 2010 8th International Conference on ICT and IEEE, **(2010)**.
[8] E. Saad, M. Awadalla, and A. Alajmi, "Dewy index based Arabic document classification with synonyms merge feature reduction", IJCSI, **(2011)**.
[9] M. L. H. Patil, and M. Atique, "A Multistage Feature Selection Model for Document Classification Using Information Gain and Rough Set".
[10] J. Ababneh, "Vector Space Models to Classify Arabic Text", International Journal of Computer Trends and Technology (IJCTT), vol. **7**, no. 4, **(2014)**, pp. 219-223.
[11] B. M. Zahran, and G. Kanaan, "Text Feature Selection using Particle Swarm Optimization Algorithm 1", **(2009)**.
[12] T. Zaki, "A Hybrid Method N-Grams-TFIDF with radial basis for indexing and classification of Arabic documents", International Journal of Software Engineering and Its Applications, vol. 8, no. 2, **(2014)**, pp. 127-144.

[13] A. A. Errub, "Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements", International Journal of Computer Applications, vol. 93, no. 6, **(2014)**.
[14] A. Alajmi, E. Saad, and M. Awadalla, "DACS Dewey index-based Arabic Document Categorization System", International Journal of Computer Applications, vol. 47, no. 23, **(2012)**, pp. 50-57.
[15] G. Forman, "An extensive empirical study of feature selection metrics for text classification", The Journal of machine learning research, vol. 3, **(2003)**, pp. 1289-1305.
[16] J. J. G. Adeva, "Automatic text classification to support systematic reviews in medicine", Expert Systems with Applications, vol. 41, no. 4, **(2014)**, pp. 1498-1508.
[17] F. Sebastiani, "Machine learning in automated text categorization", ACM computing surveys (CSUR), vol. 34, no. 1, **(2002)**, pp. 1-47.
[18] C. Zifeng, "CLDA: feature selection for text categorization based on constrained LDA", in Semantic Computing, 2007. ICSC 2007. International Conference on, IEEE, **(2007)**.
[19] Y. Xu, "A study on mutual information-based feature selection for text categorization", Journal of Computational Information Systems, vol. 3, no. 3, **(2007)**, pp. 1007-1012.
[20] F. Thabtah, "Naïve Bayesian based on Chi Square to categorize Arabic data", in proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, Citeseer, **(2009)**.
[21] K. Dave, "Study of feature selection algorithms for text-categorization, "**(2011)**.
[22] X. B. Xue and Z.-H. Zhou, "Distributional features for text categorization", Knowledge and Data Engineering, IEEE Transactions on, vol. 21, no. 3, **(2009)**, pp. 428-442.
[23] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information processing & management, vol. 24, no. 5, **(1988)**, pp. 513-523.
[24] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", Springer, **(1998)**.

