Using Some Web Content Mining Techniques for Arabic Text Classification

ZAKARIA SULIMAN ZUBI
Computer Science Department
Faculty of Science
Al-Tahadi University
P.O Box 727, Sirt, Libya
zszubi@yahoo.com

Abstract:- With the massive rise in the volume of information available on the World Wide Web these days, and the emergence requirements for a superior technique to access this information, there has been a strong resurgence of interest in web mining research. Web mining is a critical issue in data mining as well as other information process techniques to the World Wide Web to discover useful patterns. People can take advantage of these patterns to access the World Wide Web more efficiently. Web mining can be divided into three categories such as content mining, usage mining, and structure mining. In this paper we are going to apply web content mining to extract non-English knowledge from the web.

We will investigate and evaluate some common methods; using web mining systems which have to deal with issues in language-specific text processing. Arabic language-independent algorithm will be used as a machine learning system. The algorithm will use a set of features as a vector of keywords for the learning process to apply *text classification* for the system. The algorithm usually used to classify a various number of documents written in a non English text language. The techniques used in the algorithm to categorize and classified the documents are two classifiers: Classifier K-Nearest Neighbor (CK-NN) and Classifier Naïve Bayes (CNB). However, the algorithms usually depend on some phrase segmentation and extraction programs to generate a set of features or keywords to represent the retrieved web documents. A proposed Arabic text classification system will be called Arabic Text Classifier (ATC). The main goal of ATC is to compares the results between both classifiers used (CK-NN, CNB) and select the best average accuracy result rates to start a retrieving process. The theorem behind the ATC was introduced in this paper without demonstrating any practical views of the system.

Keywords: Web mining, Web content mining, Text mining, Multilingual web mining, Multilingual text mining, Data mining, Text classification, K-Nearest Neighbor, Naïve Bayes.

1 Introduction

Earlier studies of data mining have focused on structured data. However, in reality a considerable portion of the existing information are stored in text databases which consist of a large collection of documents from various sources such as books, articles, research papers, e-mail messages and web pages. With the existence of a marvelous number of these documents, it is tedious yet essential to be able to automatically systematize the documents into classes to assist document retrieval and subsequent analysis. Automatic text classification is the process of assigning a text document to one or more predefined categories based on its content [2].

There are numerous research projects examining and exploring the techniques in classifying English documents [1]. In addition to English language there are many studies in European languages such as French, German, Spanish [6], and in Asian languages such as Chinese and Japanese [27]. However, in Arabic language there is little ongoing research in automatic Arabic document classification as well. The three main consecutive phases in constructing a classification system are listed as follows:

1. Collect the text documents in corpora and tag them.

2. Select a set of features to represent the defined classes.

3. The classification algorithms must be trained and tested using the collected corpora in the first stage.

This paper attempts to achieve a better understanding of Arabic text classification techniques by using the mentioned phases. The remainder of the paper is organized as follows. Section 2 discusses related work in Arabic text classification. In section 3, expresses the foundation of web content mining. In section 4, we described the importance of text mining in web documents. In section 5, we indicate the significant of multilingual web mining. The Arabic language-independent specification was conducted in section 6. Section 7, explained in details the needs of Arabic text classification. In section 8, we present the Arabic corpora design and compilation. In section 9, we indicate all methods and algorithms used in our work. In section 10, we represent the implementation of our proposed Arabic text classifier. The results and discussions were shown in section 11. Section 12, we concludes with a summary of our work.

2 Related Work

There are many research in classifying English documents (i.e. have a survey) [20]. In addition to English language there are many researches in European languages such as German, Italian and Spanish [23], Asian languages as well, such as Chinese and Japanese [25]. However, in Arabic language there are only four researches which are: (1) El-Kourdi et mentioned in [27], use naïve bayes algorithm to automatic Arabic document classification. The average accuracy reported was about 68.78%.

The second system is called (2) Siraj from Sakhr. The system is available at (siraj.sakhr.com) but it has no technical documentation to explain the method used in the system and the accuracy of the system. The third system proposed by (3) Sawaf et. al. indicated in [30], this system uses the statistical classification methods such as maximum entropy to classify and cluster news articles, the best classification accuracy they reported was 62.7% with precision of 50% which is a very low precision in this field. In addition, (4) El-Halees indicated in [31], described a method based on association rules to classify Arabic documents. The classification accuracy reported was 74.41%.

3 Web Content Mining

Web content mining is widely used for discovering the useful information from text, image, audio or video data in the web. Web content mining occasionally is called web text mining, since the text content is the most extensively researched area. The technologies behind the use of web content mining are Natural language processing (NLP) and Information retrieval (IR).

Web content mining is mainly based on research in information retrieval and text mining, such as information extraction, text classification and clustering, and information visualization. However, it also includes some new applications, such as Web resource discovery. The main concern of this work is to focus on text mining which led us to extract non English text from web documents. Some important web content mining techniques and applications are reviewed in this paper as well.

4 Text Mining for Web Documents

As mentioned previously, text mining is often considered a sub-field of data mining and refers to the extraction of knowledge from text documents [8]. Because of a mass number of documents on the web are text documents. Text mining for web documents considered as a sub-field of web mining, or, more specifically, web content mining. Information extraction, text classification, and text clustering are patterns of text-mining applications that have been functional to web documents.

The techniques of information extraction have been applied to plain text documents; but extracting information from HTML web pages presents a different problem domain. The reason is that, HTML documents contain many markup tags that can identify useful information from the web.

On the other hand, web pages are also comparatively unstructured. Instead of a document consisting of paragraphs, a web page can be a document composed of a sidebar with navigation links, tables with textual and numerical data, capitalized sentences, and repetitive words. The range of formats and structures is very diverse across the web. If a system could parse and understand such structures, it would effectively acquire additional information for each piece of text. For example, a set of links with a heading “Link to my friends’ homepages” may indicate a set of people’s names and corresponding personal home page links. The header row of a table can also

provide additional information about the text in the table cells. Conversely, if these tags are not processed correctly but simply stripped off, the document may become much noisier.

In a previous study Chang and Lui in [7], used a PAT tree to construct automatically a set of rules for information extraction. The system, called IEPAD (Information Extraction Based on Pattern Discovery), reads an input web page and looks for repetitive HTML markup patterns. After unwanted patterns have been filtered out, each pattern is used to form an extraction rule in regular expression. IEPAD has been tested in an experiment to extract search results from different search engines and achieved a high retrieval rate and accuracy. Wang and Hu in [32], used both decision tree and SVM to learn the patterns of table layouts in HTML documents. Layout features, content type, and word group features are combined and used as a document's features. Experimental results show that both decision tree and SVM can detect tables in HTML documents with high accuracy. Borodogna and Pasi in [5], proposed a fuzzy indexing model that allows users to retrieve sections of structured documents such as HTML and XML. Doorenbos, Etzioni, and Weld in [12] also have applied machine learning in the ShopBot system to extract product information from web pages.

Some business applications are used frequently to extract useful information from web pages. For instance, FlipDog developed by the Whiz-bang labs, which helps the web to identify job openings on employer web sites. Lencom software also developed a number of products that can extract e-mail addresses and image information from the web.

Although information extraction analyzes individual web pages, text classification and text clustering analyze a set of web pages. Web pages consist mostly of HTML documents and are often noisier and less structured than traditional documents such as news articles and academic abstracts. In some applications the HTML tags are simply stripped from the web documents and traditional algorithms are then applied to perform text classification and clustering.

However, some useful characteristics of web page design would be ignored. For example, web page hyperlinks would be lost, but "Home," "Click here,'' and "Contact us," would be included as a document's features. This creates a unique problem for performing text classification and clustering of web documents because the format of HTML documents and the structure of the web provide additional information for analysis. For example, text from neighboring documents has been used in an attempt to improve classification performance. "However, experimental results show that this method does not improve performance because, often, too many neighbor terms and too many cross-linkages occur between different classes" [9, 34] .

The use of information from neighboring documents has been proposed, including the predicted category of neighbors that applies the anchor text pointing to a document and the outgoing links to other documents. It has been shown that using such additional information improves classification results [9].

Similarly, text clustering algorithms have been applied to web applications. In the Grouper system, they applied the Suffix-Tree clustering algorithm described earlier to the search results of the Husky Search system. The self-organizing map (SOM) technique also has been applied to web applications. It uses a combination of noun phrasing and SOM to cluster the search results of search agents that collect web pages by meta-searching popular search engines or performing a breadth-first search on particular web sites. On the other hand also, they used a combination of content, hyperlink structure, and co-citation analysis in web document clustering. Two web pages are considered similar if they have similar content, they point to a similar set of pages, or many other pages point to both of them.

The large amount of documents available on the web makes it an outstanding resource for linguistic studies. The digital library project groups of the University of California at Berkeley and Stanford University analyzed 88 million web pages and calculated the document frequency of the 113 million terms found in those pages. They use the web as a resource for finding phrases with high co-occurrences.

Therefore, we could conclude from all the above that text classification is on of the important approaches used in text mining, since all unsupervised documents on the web wants to categorize by a classification process. This paper will address all the related topics about text classification and proposed two classifiers for the text classification process.

5 Multilingual Web Mining

The number of non-English documents on the web continues to grow-more than 30 percent of web pages are in a language other than English. In order to extract a non-English knowledge from the web, web mining systems have to deal with issues in language-specific text processing.

We won't consider that as a problem because the essential algorithms behind most machine learning systems are language independent. The majorities of algorithms, such as text classification and clustering, need only a set of features (a vector of keywords) for the learning process. However, the algorithms regularly depend on some phrase segmentation and extraction programs to generate a set of features or keywords to represent web documents. Many existing extraction programs, especially those employing a linguistic approach, are language-dependent and work only with English texts. In order to perform analysis on non-English documents, web mining systems must use the corresponding phrase extraction program for each language.

Other learning algorithms, such as information extraction and entity extraction, also have to be tailored for different languages. Some segmentation and extraction programs are language-independent. These programs usually employ a statistical or a machine learning approach. Most common approaches to categorize and classified the documents are Key Nearest Neighbor (K-NN) and the Naïve Bayes.

These algorithms were tested on Chinese documents as a language-independent technique for key phrase extraction and have shown promising results in [7]. These algorithms do not rely on specific linguistic rules; they can be easily modified to work with different languages as well.

In this paper we tried to indicate the must relevant text classification algorithms for Arabic text. These algorithms will be evaluated on a data set in an Arabic corpus.

6 Arabic Language-Independent Specification

A language-independent specification is a programming language specification which provides a common interface functional for defining semantics applicable headed for arbitrary language bindings; in other words, language-independent specifications are language-agnostic [13]. It is also mitigate the risk that a certain language binding might reduce compatibility with other languages; an ideal language-independent specification allows the language bindings to take advantage of features of a programming language uncompromisingly. The use of language-independent specification guides us to use language-independent rules wherever possible, i.e. use rules that can be applied to any new language that will be added to the system.

It keeps the language- independent specific resources and tools to a minimum: such tools are, for instance, linguistic software like part-of-speech taggers and parsers. Language independent specific resources are morphological or subject domain-specific dictionaries, linguistic grammar rules, etc. As acquiring or developing such language-independent specific resources is difficult and expensive and that is why we try to avoid developing such a language to reduce complexity.

It carries on the application modular by storing necessary language-specific resources outside the rules, inside a language-specific parameter files. It gives the application more capabilities to add new languages to be plugged in the system more easily.

For the language-specific information that cannot be done without, use bottom-up, data-driven bootstrapping methods to create the monolingual resources.

It avoids the language pair-specific resources and procedures because almost exponential growth of language pairs would automatically limit the number of languages a system can deal with.

7 Arabic Text Classification

Arabic is the mother language of more than 300 million people [16]. Unlike Latin-based alphabets, the direction of writing in Arabic is from right to left; the Arabic alphabet consists of 28 letters. Arabic words have two genders, feminine and masculine; three numbers, singular, dual, and plural; and three grammatical cases, nominative, accusative, and genitive [4]. A noun has the nominative case when it is subject; accusative when it is the object of a verb; and the genitive when it is the object of a preposition. Words are classified into three main parts of speech, nouns (including adjectives and adverbs), verbs, and particles.

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 highly inflectional and derivational language which makes monophonical analysis a very complex task [21]. Also, in Arabic scripts some of the vowels are represented by diacritics that usually left out in the text which create ambiguity in that text. In addition, Arabic scripts do not use capitalization for proper nouns which are necessary in classifying documents classification [21].

Much of the work in text classification takes care of documents as a bag-of-words with the text represented as a vector of a weighted frequency for each of the individual words or tokens. Such a simplified illustration of text has been shown to be quite effective for a number of applications [11] , [31]. There are several efforts to improve text illustration using concepts or multi-word terms [25].

El-Kourdi et al in [16], used Naïve Bayes algorithm to automatically classify Arabic documents. The average accuracy reported was about 68.78% [16]. In [30] they used statistical classification methods such as maximum entropy to classify and cluster news articles. The best classification accuracy they reported was 62.7%. In addition, [15] described a method based on association rules to classify Arabic documents. The classification accuracy reported was 74.41%. Al-Fedaghi and Al-Anzi's algorithm tries to find the root of the word by matching the word with all possible patterns with all possible affixes attached to it [14].

The morphology system uses different algorithms to find the roots and pattern [3]. This algorithm removes the longest possible prefix, and then extracts the root by checking the first five letters of the word. This algorithm is based on an assumption that the root must appear in the first five letters of the word. Khoja in [16], [23], has developed an algorithm that removes prefixes and suffixes, all the time checking that it's not removing part of the root and then matches the remaining word against the patterns of the same length to extract the root.

In this paper we are going to evaluate some recent average accuracy reports from researches that use the same classifiers we used in our proposed system.

8 Arabic Corpora

A corpus of Arabic text documents was collected from online Arabic newspapers archives, including Al-Jazeera, Al-Nahar, Al-hayat, Al-Ahram, and Al-Dostor as well as a few other specialized websites.

Since the Arabic corpus was one the difficult issues that came across this work when evaluating text categorization algorithms, for this a small corpus was proposed and prepared which consists of 1562 documents belonging to 6 different categories. Another difficulty was the Arabic morphology and text classification depends on the contents of documents, a huge number of features or keywords can be found in Arabic text such as morphemes that may generated from one root which may tend to a poor performance in terms of both accuracy and time. To deal with this problem, the focus was about unvowelized keywords in this case the numbers of extracted features were reduced. All documents, whether training documents or documents to be classified went through a preprocessing phase removing punctuation marks, stop words, diacritics, and non letters.

9 Methods and Algorithms

Text classification theory has been concerned during the last 25 years, a huge number of algorithms have been initiated and evaluated in the last three decades, some surveys was introduced as well in [19]. These surveys address the most common text classification algorithms and a lot of works have been accomplished to evaluate and compare between these algorithms.

The text classification problem is composed of several sub problems, which have been deliberated intensively in the literature such as the document indexing, the weighting assignment, the document clustering, the dimensionality reduction, the threshold determination and the type of classifiers. Document indexing is related to with the way of extracting the document's keywords, two main approaches to achieve the document indexing, the first approach considers index terms as bags of words [29], and the second approach regards the index terms as phrases [26,24]. A drawback of the first approach is that it complicates the extraction process of index term by increasing the number of words that must be dealt with in the document as well as dealing with unrelated words to any category. Classifying Arabic documents requires accomplishing some preprocessing steps for the documents through stemming the words; this

process is quite a major issue in terms of reducing the number of related words in a document. Several techniques have been established to perform such preprocessing tasks such as stemming, root extraction and thesaurus.

Weight assignment techniques associate a real number that ranges from 0 to 1 for all documents' terms in [19], the weights will be required to classify new arrived documents. Different information retrieval models use different methodologies to compute these weights, for example the Boolean model assigns either 0 or 1 for each index term. In contrast, vector space model compute tf-idf factor shown in [26], which ranges from 0 to 1, this model is further described in the following section.

Two main divisions for learning based text classification algorithms, the inductive learning algorithms and the clustering-based algorithms. A text classification algorithm used different decision tree models to classify documents through building a tree by computing the entropy function of the selected index terms shown in [18,4] such as ID3 in [26] and C4.5 in [17,33]. Another inductive learning algorithm based on probabilistic theory, different models were emphasized such as Naive Bayesian models [23,35,36] which have depicted good results in the text classification field. Historically, the most widely famous Naïve Bayesian model was known as the binary independent classifier introduced in [28]. One of the first attempts to address clustering techniques for text classification problem was stated through introducing a comparison between the k-means algorithms and hierarchical clustering algorithms in [17].

The conducted results showed better performance for the hierarchical algorithms although they were slower than the k-means algorithm. Different text classification methods have been emerged to categorize documents such as K-Nearest Neighbor KNN shown in [35], the K-NN various models compute the distances between the document index terms and the known terms of each category by applying distance functions such as cosine, dice similarity or Euclidian functions, the returned classes are the kth classes with highest scores.

The accuracy will be tested by K-fold cross-validation method. In K-fold cross-validation, the original sample is partitioned into K sub samples. Of the K sub samples, a single sub sample is retained as the validation data for testing the model, and the remaining $K-1$ sub samples are used as training data. The cross-validation process is then repeated K times (the folds), with each of the K sub samples used exactly once as the validation data. The K results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used.

In stratified K-fold cross-validation, the folds are selected so that the mean response value is approximately equal in all the folds. In the case of a dichotomous classification, this means that each fold contains roughly the same proportions of the two types of class labels.

9.1 Preprocessing phases

The data used in this work are collected from many Arabic sites. The data set consist of 1562 Arabic documents of different lengths that belongs to 6 categories, the categories are (Economic "اقتصادية", Cultural "ثقافة" , Political "سياسية" , Social "اجتماعية" , Sports "رياضة " , General "عامة"), Table 1 represent the number of documents for each category.

Category Name	Number of Documents
Cultural News	258
Sports News	255
Economic News	250
Social News	258
Political News	250
General News	255
Total	1562

Table 1: Number of Documents per Category

Standardized text classification process passes several major phases, the *first phase* is the preprocessing step where documents are prepared to make it adequate for further use, stop words removal and rearrange of the document contents are some steps in this phase.

The *second phase* is the weighting assignment phase, it is defined as the assignment of real number that relies between 0 and 1 to each keyword and this number indicates the imperativeness of the keyword inside the document. Different methods have been developed and the most widely used model is the

tf-idf weighting factor in [26]. This weight of each keyword is computed by multiplying the term factor (*tf*) with the inverse document factor (*idf*) where:

- F_{ik} = Occurrences of term t_K in document D_i.
- tf_{ik} = fik/max (f_{il}) normalized term frequency occurred in document.
- dfk = documents which contain t_k .
- idfk= log (d/d_{fk}) where d is the total number of documents and dfk is number of document s that contains term t_k.
- wik = tfik * idfk for term weight, the computed w_{ik} is a real number □[0,1].

Formally, clustering is defined as the process of grouping a set of physical or abstract objects into classes of similar objects [28, 22]. Similarity between objects can be measured in different ways, the most widespread clustering techniques measure the distances between the arrived objects and an arbitrary selected point known as the center of the cluster.

Distance functions calculate the distance between the set of extracted features of the new object and the center's features. Euclidian function formula in [36,22] is in the form of: Another form of measurements tries to find the closeness criterion between documents through calculating the similarity between the key words of such documents. The most famous measurement is the cosine similarity measurement.

9.2 Classifier Naive Bayesian (CNB)

Bayesian learning is a probability-driven algorithm based on Bayes probability theorem as the follow:

$$p(A|B) = \frac{p(A)\,p(B|A)}{p(B)}$$

This function is highly recommended in Arabic text classification it helps us to compute the conditional probability based on the reverse relation. The function generally can be conducted to our problem as follow:

$$p(class|document) = \frac{p(class)\,p(document|class)}{p(document)}$$

P(class| document) : It's the probability of *class* given a document, or the probability that a given *document D* belongs to a given *class C*, and that is our target.

P(document) : The probability of a document, we can notice that p*(document*) is a Constance divider to every calculation, so we can ignore it.

P(class): The probability of a class (or category), we can compute it from the number of documents in the category divided by documents number in all categories.

P(document | class) : It's the probability of document given class, and documents can be modeled as sets of words thus the *P(document | class)* can be written like that:

$$p(document|class) = \prod_i p(word_i\,|class)$$

So:

$$p(class|document) = p(class)\prod_i p(word_i\,|class)$$

$p(word_i\,|C\,)$: Probability that the i-th word of a given document occurs in a document from class C, and this can be calculated as the follow:

$$p(word_i|C) = \frac{times\ te\ word\ occurs\ in\ ta\ category\ C + 1}{number\ of\ words\ in\ C\ category + size\ of\ te\ vocabulary\ table}$$

The Naive Bayesian can often outperform more sophisticated classification methods. Classifier task is to categorize incoming objects to their appropriate class [28].

In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability. Finally, the new object X is classified based on the higher posterior probability [28]. The CNB can handle a random number of independent variables whether continuous or categorical.

The CNB works as follows:

Each data sample is represented by an n-dimensional

feature vector, $X = (x_1, x_2,...x_n)$, representing n measurements made on the sample from n attribute, respectively, $A_1, A_2,...,A_n$.

Suppose that there are m classes. $C_1, C_2,...C_m$. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditional on X. That is, the naïve Bayesian classifier assigns an unknown sample X to the class C_i if and only if:

$P(C_i|X) \square P(C_j|X)$, where $1 \leq j \leq m$, and $j \neq i$.

The class C_i for which $P(C_i\backslash X)$ is called as the maximum posteriori hypothesis. By the Bayesian theorem :

$$P(C_i \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)}$$

As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class prior probabilities are unknown, then it is commonly assumed that the classes are equally likely, that is:

$$P(C_1) = P(C_2) = ... = P(C_m)$$

Note that the Class prior probabilities may be estimated by

$P(C_i) = \frac{si}{s}$,where , si is the number of training samples of class C_i and s is the total number of training samples.

In order to reduce computation in evaluating, the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample that is there are no dependence relationships among the attributes. Thus,

In order to classify unknown samples X, $P(X\backslash C_i)P(C_i)$ is evaluated for each class Ci. Sample X is then assigned to the class C_i if and only if.

9. 3 Classifier K-Nearest Neighbor (CK-NN)

K-Nearest Neighbor is a widely used text classifier especially in non English text mining because of its simplicity and efficiency. Its training-phase consists of nothing more than storing all training examples as classifier, thus it has often been called as lazy learner since it postpones the decision on how to generalize beyond the training data until each new query instance is came across.

The CK-NN is a supervised learning algorithm where the result of a new occurrence query is classified based on the K-nearest neighbor category measurement.

The idea of this algorithm is to classify a new object based on attributes and training samples [10]. The nearest neighbor classifier is based on learning by likeness. The training samples are described by n-dimensional numeric attributes. Each sample represents a point in an n-dimensional pattern space. In this way, all of the training samples are stored in an n-dimensional pattern space [35].

We will introduce an example which demonstrates the general aspects of this algorithm in detail. The K nearest neighbor algorithm is a very simple algorithm in particular.

It works based on minimum distance from the query instance to the training samples to determine the K nearest neighbors. After we gather K nearest neighbors, we take simple majority of these K-nearest neighbors to be the prediction of the query-instance [33].

The data for CK-NN algorithm consists of several multivariate attributes names Xi that will be used to classify the object Y. The data of CK-NN can have any measurement scale from ordinal, nominal, to quantitative scale, this paper deals only with quantitative Xi and binary (nominal) Y [33].

Suppose that the K factor is set to be equal to 8 (there are 8 nearest neighbors) as a parameter of this algorithm. Then the distance between the query-instance and all the training samples is computed [7].

Because there are only quantitative X_i, the next step is to find the K-nearest neighbors. All training samples are included as nearest neighbors if the distance of this training sample to the query is less than or equal to the K^{th} smallest distance. In other words, the distances are sorted of all training samples to the query and determine the K^{th} as a minimum distance described in [33]. The unknown

sample is assigned the most common class among its k nearest neighbors [20].

As illustrated above, it is necessary to find the distances between the query and all training samples.

These K training samples are the closest K nearest neighbors for the unknown sample. Closeness is defined in terms of Euclidean distance, where the Euclidean between two points, $X = (x_1, x_2,...,x_n)$ and $Y = (y_1, y_2,...,y_n)$ is:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

10 Implementation of the ATC

The implementation of the proposed Arabic Text Classifier (ATC) demonstrates the importance of classifying the Arabic text on the web documents needs for information retrieval to illustrate both keywords extraction and text classifiers which will be described in the following subsections:

Keywords extraction: Text web documents are scanned to find the keywords each one is normalized. Normalization process consists of removing stop words, removing punctuation mark in Arabic letters such as the normalization of the (hamza (ﺇ) or (ﺃ) or (ﺁ)) in all its forms to (alef (ﺍ)) only, the ى replaced by ي , and the ة to ه, moreover, remove any Maad "-" character and finding keywords root or stem. In this work light stemmer is used to find the requested keywords. Table 2. Shows examples of the stop words.

Stop word (Arabic)	English
في	in
على	on
أين	where

Table1: The Example of stop words.

Terms weighting: there are many approaches used to index terms but all of them shares the same characteristics. One criterion is the more number of times a term occurs in documents which belongs to some category, the more it is relative to that category. The second criterion is the more the term appears in different documents representing different categories; the less the term is useful for discriminating between documents as belonging to different categories.

The most commonly used weighting approach is the tf×idf, in our implementation we used the Normalized tf×idf to overcome the problem of variant documents lengths.

Algorithms implementation: As we declared previously our proposed implementation is mainly developed for testing the effectiveness of CK-NN and CNB algorithms when applied to Arabic text. A set of labeled text documents are supplied to the system, the labels are used to indicate the class or classes that the text document belongs to. All documents belonging to the data set should be labeled in order to learn the system and then test it.

The system distinguishes the labels of the training documents but not those of the test set (i.e., ignores them). The system classifies a test document comparing it to all the examples it has (i.e., the training set), the comparison is done using a two previous classifiers mentioned earlier. In this paper we are going to estimate some recent average accuracy reports from researches that use the same classifiers we used in our proposed system, with our ATC system.

The system will compare these results and select the best average accuracy result rates for each classifier and uses the greater average accuracy result rates in the ATC system. The system will choose the higher rate to start the retrieving process.

11 Results and Discussion

In the evaluation process of our system, we have used the data set described earlier; the documents belong to 6 different categories. The documents of the set were used to select keywords suitable to represent the categories. To test the system, the documents in the data set were preprocessed to find main categories.

Various splitting percentages were used to see how the number of training documents impacts the classification effectiveness. We also used different k values starting from 1 and up to 20 in order to find the best results for CK-NN. Effectiveness started to decline at k>15.

The proposed system has been built empirically to make comparison between the two algorithms and make labeling to the sample data, the classifier has been indicated in the system also depending on some training data. Preprocessing on the data has been done like Cleaning (releasing stop words), words division.

The *k-fold cross-validation* method is used to test the accuracy of the system. Here we will demonstrate some results from related work. The below table illustrates some results to some recent works, that have applied the two algorithms with these results.

The results of the conducted experiments are included on the last column in Table 3. Our result is roughly near from the other developer's results. It is induced from the below results that the Classifier K-Nearest Neighbors (CK-NN) with an average (86.02%) has better than Classifier Naïve Bayesian that had (77.03%). It means that the ATC system in this case will use the CK-NN for Arabic text classification and extraction instead of CNB. But it never means that the ATC system won't use CNB as an Arabic text classification. Because some times in case of Multilanguage text the CNB shows a higher average accuracy rates, more than other classification algorithm.

N.O	Researchers	El-Kourdi et al.	*Siraj*	*Sawaf*	*El-Halees*	Our result (ATC)
1	Training	8560	8560	8560	6740	1562
2	Test	2780	2780	2780	2605	798
3	Topics	93	88	90	90	6
4	CNB	68.78%	72.02%	50%	74.41%.	77.3 %
5	CK-NN	75.02%	82.03%	62.7%	85.01%	86.2%

Table 3. The results of our experiments.

12 Conclusion

In this paper an evaluation to the use of Classifier K-Nearest Neighbor (CK-NN) and Classifier Naïve Bayes (CNB) to the classification Arabic text was considered. We developed a special corpus consists of 1562 documents that belong to 6 categories. We have also extract feature set of keywords and terms weighting in order to improve the performance.

The result in the previous section demonstrates that two algorithms were applied to the Arabic text; a satisfactory number of patterns for each category were indicated. The selection of the feature space and the training data set used was also included. The accuracy was measured by the use of k-fold cross-validation method to test the accuracy of the system. The value of k can extremely influence the accuracy of any text classification algorithm.

Finally, we proposed an empirical Arabic text classifier system called ATC. The system compares the results between both classifiers used (CK-NN, CNB) and select the best average accuracy result rates.

References:

[1] Aas K. and Eikvil L. (1999). Text Categorisation: A survey. Technical report, Norwegian Computing Center.

[2] Alexandrov M., Gelbukh A. and Lozovo. (2001). Chi-square Classifier for Document Categorization. 2nd International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.

[3] Al-Shalabi R. and Evens M. (1998). A Computational Morphology System for Arabic. Workshop on Semitic Language Processing. COLING-ACL'98, University of Montreal, Montreal, PQ, Canada.pp. 66-72.

[4] Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone, 1984. Classification and Regression Trees. 1st Edn., Chapman Hall, New York, ISBN: 0-412-04841-8.

[5] Bordogna, G. Pasi, G. , A user-adaptive indexing model of structured documents, Fuzzy Systems, 2001. The 10th IEEE International Conference on 2-5 Dec. 2001, Volume: 2, On page(s): 984- 989 vol.3, ISBN: 0-7803-7293-X.

[6] Ciravegna F., Gilardoni L., Lavelli A., Ferraro M., Mana N., Mazza, S., Matiasek J., Black W. and Rinaldi F. (2000). Flexible Text Classification for Financial Applications: the FACILE System. In Proceedings of PAIS-2000, Prestigious Applications of Intelligent Systems sub-conference of ECAI2000.

[7] Chang, C. and Lui, S. 2001. IEPAD: information extraction based on pattern discovery. In Proceedings of the 10th international Conference on World Wide Web (Hong Kong, Hong Kong, May 01 - 05, 2001). WWW '01. ACM, New York, NY, 681-688. DOI= http://doi.acm.org/10.1145/371920.372182

[8] Chen, K. H. and Chen, H. H. 2001. The Chinese text retrieval tasks of NTCIR workshop 2, In Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization.

[9] Chakrabarti, B. Dom, P. Indyk, New block Enhanced hypertext categorization using hyperlinks. In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 307--318, Seattle, Washington, 1998.

[10] Danesh, A., B. Moshiri and O. Fatemi, 2007. Improve text classification accuracy based on classifier fusion methods. Proceeding of the 10th International Conference on Information Fusion, July 9-12, IEEE Computer Society, USA., pp: 1-6. Doi: 10.1109/ICIF.2007.4408196.

[11] Diederich J., Kindermann J. L., Leopold E. and Paaß G. (2003). Authorship attribution with support vector machines. Applied Intelligence, 19(1/2): 109-123.

[12] Doorenbos, R. B., Etzioni, O., and Weld, D. S. 1997. Learning to Understand Information on the Internet: AnExample-Based Approach. J. Intell. Inf. Syst. 8, 2 (Mar. 1997), 133-153.

[13] Dunham M. H. (2003). Data Mining: Introductory and Advanced Topics. Prentice Hall. AUTOMATIC ARABIC TEXT CLASSIFICATION 83 JADT 2008

[14] Duwairi R. M. (2005). A Distance-based Classifier for Arabic Text Categorization. In Proceedings of the International Conference on Data Mining, Las Vegas USA.

[15] El-Halees A. (2006). Mining Arabic Association Rules for Text Classification. In Proceedings of the first international conference on Mathematical Sciences. Al-Azhar University of Gaza, Palestine.

[16] El-Kourdi M., Bensaid A. and Rachidi T. (2004). Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. 20th International Conference on Computational Linguistics. August, Geneva.

[17] Evegeny, G. and M. Shaul, 2004. Text classification with support vector machine learning with many relevant features. Proceedings of the 21st International Conference Machine Language, 2004, ACM Publising, pp: 1-8. DOI: 10.1007/BFb0026683.

[18] Evegniy, G. and M. Shaul, 2004. Text Classification with many redundant features: Using aggressive feature selection to make svms competitive with C4.5. Proceeding of the 21st International Conference Machine Learning, July 4-8, Banff, Alberta, Canada, pp: 41.

[19] Fabrizio, S., 2002. Machine learning in automated text classification. ACM Comput. Surveys, 34: 1-47. DOI: 10.1145/505282.505283.

[20] Forman G. (2003). An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3, 1289-1305.

[21] Hammo, B., Abu-Salem, H., Lytinen, S., Evens, M., QARAB: A Question Answering System to Support the Arabic Language. Workshop on Computational Approaches to Semitic Languages. ACL 2002, Philadelphia, PA, July. p 55-65 (2002).

[22] Jiawei and K. Micheline, 2001. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publishers, San Francisco, ISBN: 1-55860-901-6.

[23] Larkey L. and Connell M. E. (2001). Arabic information retrieval at UMass in TREC-10. In Proceedings of TREC, Gaithersburg: NIST.

[24] Maria, F.C., M. Stan and S. Fabrizio, 2001. A Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Classification. In: Text Databases and Doc. Management Theory and Practice, Chined, A.G. (Ed.). IGI Publishing, Hershery, PA., USA., ISBN: 1-878289-93-4, pp: 78-102.

[25] Mesleh A. A. (2007). Chi Square Feature Extraction Based Svms Arabic Language Text

Categorization System. Journal of Computer Science 3(6): 430-435.

[26] Norbert, F., H. Stephen, L. Gerhard, S. Michael and T. Kostandinos, 1991. AIR/X: A rule-based multistage indexing system for large subject fields. Proceedings of RAIO~91, the 3rd International Conference Rech. D'Information Assiste par Ordinateur, 1991, Barcelona, Spain, pp: 606-623.

[27] Peng F., Huang X., Schuurmans D. and Wang S. (2003). Text Classification in Asian Languages without Word Segmentation. In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL 2003), Association for Computational Linguistics, Sapporo, Japan.

[28] Richard, D. and J. Anil, 1988. Algorithms for Clustering Data. 2nd Edn., Prentice Hall, USA., ISBN: 0-13-022278-X.

[29] Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. J. Inform. Process. Manage., 24: 513-523.

[30] Sawaf H., Zaplo J. and Ney H. (2001). Statistical classification methods for Arabic news articles. Natural Language Processing in ACL2001, Toulouse, France.

[31] Sebastiani F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, Vol. 34 number 1. pp.1-47.

[32] Wang, Yalin, J. Hu, (2002), "A machine learning based approach for table detection on the web", in Proceedings of the Eleventh International World Wide Web Conference (WWW2002), pp. 242-250, Hawaii, USA.

[33] William, C. and H. Haym, 1998. Joins that generalize text classification using WHIRL. Proceedings of 4th International Conference Knowledge Discovery and Data Mining, 1998,CiteSeerX Press, pp: 169-173.

[34] Y. Yang, S. Slattery, and R. Ghani, A study of approaches to hypertext categorization, Journal of Intelligent Information Systems, 219--241, 2002.

[35] Yiming, Y. and L. Xin, 1999. A re-examined of text classification method. In Proceedings of SIGIR~99 Conference, pp: 42-49.

[36] Yonghong, L. and A.K. Jain, 1998. Classification of text documents. Proceedings of the 14th International Conference on Pattern Recognition, August 16-20, Brisbane, pp: 1295-1297. DOI: 10.1109/ICPR.1998.711938.