A New Machine Learning Approach for Arabic/English Documents Classification

Walid Mohamed Aly[1], Wafaa Hanna Sharaby[2] and Hany Atef Kelleny[3]

[1] College of Computing and Information Technology,
Arab Academy for Science, Technology & Maritime Transport
Alexandria, Abu Qir, Egypt

[2] Departments of Computer Science and Information Systems,
Higher Institute for Specialized Technological Studies, Future Academy
Cairo, Misr-Ismailia Road, Egypt

[3] College of Computing and Information Technology,
Arab Academy for Science, Technology & Maritime Transport
Cairo, Masakin Sheraton, Egypt

Abstract

This paper aims at developing a system that is capable of classifying Arabic and English un-structured documents; it proposes to classify these documents in consecutive two phases. In the first phase, incremental Automated Domain-Meta-Document Construction (ADC) algorithm is applied as a new automated machine learning approach. ADC constructs updatable summarized Domain-Meta-Documents, which corresponds to the trained classified documents. The output would be stored in a knowledge base in order to help in the classification process.

In the second phase, an enhanced supervised classification algorithm based on automated calculation of threshold value would utilize the previously generated Domain-Meta-Documents to classify the incoming Dataset. To evaluate the performance of this proposed approach, two experiments were conducted on two standard dataset, namely Corpus of Contemporary Arabic (CCA) and Newsgroup 20, whose results revealed that the proposed classification approach outperformed the compared classification algorithms (C4.5 and Back Propagation Neural Network) in different measures. The general accuracy of the proposed system was found to be about 95%.

Keywords: *Unstructured Documents, Machine Learning, Classification, Threshold.*

1. Introduction

Researches have found that today's organizations are faced with exponential growth of unstructured documents coming from different sources and in a variety of formats.

IBM estimated that 70% to 80% of their company information is currently held in unstructured forms [1]. EMC Corporation estimated that unstructured content is growing between 65% and 200% per annum [2]. The enlargement of unstructured content in organizations is estimated to be growing at a rate of 800 MB per person per year [2], which called for researches on Documents Classification (DC) to classify these unstructured documents seeking efficient information retrieval. The main problem with these unstructured documents is that they neither have indexes, nor keywords, or fixed fields to be used in classifications and retrievals. This paper presents a new methodology for unstructured Arabic or English documents classification in which automated Domain-Meta-Documents are constructed. In an ensuing stage, the classifier operates a number of processes on the unclassified datasets to classify them.

The rest of this paper is organized as follows: Section 2 presents a thorough description and background of Machine Learning techniques, Section 3 presents a review of related works, and section 4 introduces the proposed system and section 5 shows the Results and evaluation, finally, the last section gives a concluding remark.

2. Background

2.1 Machine Learning

Machine Learning [ML] is defined as "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [3].

There are three classes of ML techniques, namely, supervised, unsupervised, and semi-supervised learning techniques. The supervised ML technique induces a function from labeled training data, whereas, the

unsupervised technique attempts to find concealed structure in unlabeled data which poses a problem in this learning technique, finally, semi-supervised ML technique employs both labeled and unlabeled data for training.

ML for DC is concerned with finding the best model for a classifier after being trained using labeled data. Many of ML approaches for DC were examined in various Arabic/English DC researches. These approaches include decision trees, artificial neural networks and support vector machines.

2.2 Documents Classification

DC is a multi-phase process in which documents are assigned to one or set of predefined categories based on their content of keywords. DC is a supervised learning process which involves tokenization, pre-processing, indexing, constructing a classifier, and finally classifying the documents.

Documents are initially tokenized by performing a lexical analysis to subset them into a list of tokens. This list is generated by partitioning input stream of characters into a stream of words or tokens by employing white space or punctuation marks, yielding a list of tokens. This phase is subsequently followed by some preprocessing techniques such as removing Stop-words or negative dictionary as in "in", "the", "of"… etc., in other words, eliminating them so as not to be selected as index terms. This list is used to prepare the vectors of data to be efficiently trained by the classifier. Then in the indexing phase, terms are selected and weighted and an internal representation of the documents is created. This is followed by the construction of a classifier using a learning technique. In this learning phase, the classifier is trained with a set of classified documents to enable it to identify and classify the incoming documents. This classifier receives the testing data subsequently as input to pursue the classification process. Ultimately, the output of this process is evaluated using various measures, such as recall, precision, and F-measure.

3. Review of Literature

On examining the existing researches related to unstructured documents classification, it is lucid that there are various methods that have been developed and utilized in various contexts and studies, especially ML techniques that are explored in a wide range of studies.

Rule based classification algorithms, comprehensible to decision makers, is the easiness of the output returned rules which may even be manually updated. C4.5 classification algorithm applies decision tree approach in constructing the classification model. Using a rule induction approach, RIPPER produces rules rapaciously. PART, a hybrid learning approach, associates both of C4.5 and RIPPER to build the classifier. OneRule (1R), a simple procedure, generates one level decision tree to extract the rules from the data set.

The performance of four different Rule Based classification data mining techniques (C4.5, RIPPER, PART, & OneRule) was investigated in [4]. Given results indicated that C4.5 was found to be the most applicable classification algorithm in which it derived higher results in all evaluation criteria than RIPPER, and PART, whereas OneRule was the least applicable classification algorithm. It must be noted that RIPPER, C4.5 and PART had similar performance results with reference to error rate.

Arabic text categorization is investigated using OneRule, rule induction RIPPER, decision trees (C4.5), and hybrid PART Rule-Based classification approaches in data mining in [5]. The results signified that the hybrid approach of PART outperformed the rest of the algorithms. One Rule algorithm was found to be the least suitable classification method owing to the low rates of recall and precision.

Stemming algorithm was applied to pre-process Arabic words and extract the keywords in Arabic DC system in [6]. This algorithm was designed to remove the most common affixes from the word together with the insignificant words. Achieved results illustrated that the average performance between applying the stemming and without these stemming techniques was 0.952 and 0.808, respectively.

Topic Analyzer, an automatic technique, was operated for classifying large volumes of Arabic dataset with no prior training data and no classification scheme in [7]. This system integrated with different features extraction, selection and classification methods which enabled it to accommodate any textual data. Achieved results showed that the average F-measure was 85.8%.

Decision Trees algorithm (ID3), one of the ML techniques, is employed in classifying Arabic documents in [8]. Results demonstrated that on the application of a hybrid approach of Document Frequency (DF) threshold using an embedded information gain criterion of the decision tree algorithm, feature selection criterion, the effectiveness of the improved classifier increased the general accuracy to be about 0.93.

The Boosting technique and its effectiveness with the classification of Arabic documents were being explored in [9]. A Term Space Reduction process was operated, applying two very well known feature selection techniques: Information Gain, Chi-Square. A comparison among AdaBoost.M1 with C4.5, Naïve Bayesian (NB), Naïve Bayes Multinomial (NBM), and SVM algorithms had been conducted and results revealed that SVM and NBM were the best performing algorithms, whereas the AdaBoost.M1 had significantly enhanced the performance of C4.5.

Artificial Neural Networking (ANN) for classifying Arabic language documents using Singular Value Decomposition (SVD) method is introduced in [10], in which ANN combined with SVD achieved 88.33% accuracy compared to the basic ANN, which yielded 85.75%. The significance of the results is manifested in the fact that ANN model using SVD method is more capable for capturing the non-linear relationships between the input document vectors and the document categories than that of the basic ANN model.

The Back-Propagation Neural Network (BPNN) training algorithm is utilized in the process of classifying Arabic text using five dimensionality reduction techniques is explored in [11]. Stemming, Light-Stemming, Document Frequency (DF), Term Frequency Inverse Document Frequency (TFIDF) and Latent Semantic Indexing (LSI) methods were found for dimensionality reduction. Obtained results demonstrated that Back-Propagation learning in NNs was able to give good categorization performance.

Finally, from all of these studies we concluded that the need to develop more improved systems to dealing with unstructured documents became absolutely essential to improve the results and increase the quality of the documents classification process.

4. The Proposed Improved Classification System of Arabic/English Documents

The proposed system aims at classifying unstructured documents by utilizing two successive phases in order to classify the received Arabic/English unstructured documents with high efficiency and accuracy.

4.1 The Proposed System Structure

The structure of the proposed system is shown in Fig.1. This system comprises of two consecutive phases which illustrate the details of the process.

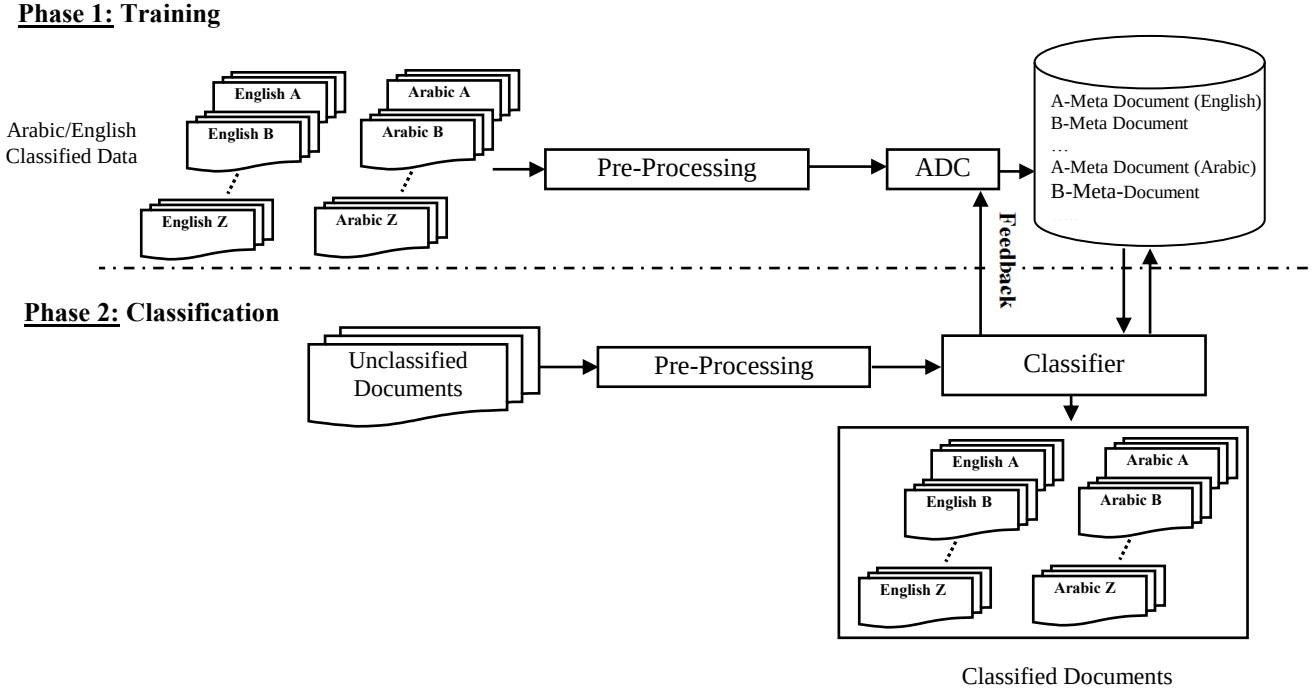

Fig. 1. Proposed system structure.

The first Phase aims at constructing automated Domain-Meta-Documents that correspond to the input set of Arabic/English classified documents, the employed system will explore and tokenize these documents then remove stop words, stem words using pre-processing techniques, which would be subsequently be processed by the ADC algorithm to ultimately generate the targeted Domain-Meta-Documents.

In the second phase, a set of unclassified Arabic or English documents is pre-processed by operating stop-words removal and stemming yielding processed documents, which would be calculated via normalized term weighting technique together with the previously generated Domain-Meta-Documents.

This phase essentially employs the weighted Arabic or English dataset and the generated Domain-Meta-Documents, in which it classifies the unstructured Arabic or English datasets using a classifier algorithm yielding classified documents. This algorithm assigns each input unclassified document to its corresponding class based on the threshold degree of similarity between the document and the Automated Domain-Meta-Document.

As shown in Fig. 1, in case there is a document that does not belong to any of the predefined set of classes by not achieving the required similarity threshold value, the classifier sends feedback to the ADC algorithm to create a new Domain-Meta-Document entry to this unclassified document and assign it to unlabeled cluster.

4.2 Data Pre-Processing

The proposed system applied the following pre-processing techniques as a stepping-stone sin the classification process. Initially, the incoming data are partitioned into a list of tokens, and then this data would be represented by a Vector Space Model (VSM) as one vector in a vector space. Then, Stop-Words as for example, "من", "الى", "هو"… etc. would be removed in the case of Arabic language whereas in English stop-words as "the", "a", "and"… etc. would be pulled out. Subsequently, to reduce the morphological disparity of the words to their stem or root, the Shereen Khoja's stemming algorithm would be employed to improve the process of automatic classification for Arabic language documents, whereas for English documents, the proposed system will use the Porter stemming algorithm [12] to improve the process of automatic indexing by reducing the morphological variants. The Khoja's stemmer is one of the superior open source root-based stemmers, which was selected scrupulously based on a comparative study that was conducted among three Arabic morphological analyzers and stemmers, namely, Khoja's stemmer, Buckwalter analyzer, and Tri-literal Root Extraction Algorithm; in which Khoja's stemmer was found to achieve the highest accuracy results [13]. In the case of English documents, the proposed system will use The Porter's stemming algorithm.

4.3 A Normalized Term Weighting Equation

Term Frequency Inverse Document Frequency (TF_IDF) is the common method of determining the weights of the term. This method depends on the frequency of the term within a document and in the entire set of documents together. Based on this method the proposed system uses the normalized term weighting Eq. (1) [14]:

$$w_{ik} = \frac{tf_{ik} \log(\frac{N}{n_k} + \alpha)}{\sqrt{\sum_{k=1}^{n} (tf_{ik})^2 \times \log^2(\frac{N}{n_k} + \alpha)}} \qquad (1)$$

Where tf_{ik} is the number of occurrences of term i in the document k; and N is the total number of documents in the collection; and n is the total number of documents in the collection which has the term k.

The α constant added to the formula because the term may appear in all documents, in this case during the calculation of IDF, the log (N/N) = log (1) = 0. The value used for α constant is 0.01.

4.4 Automatic Domain-Meta-Document Construction Algorithm

To construct an Automated Domain-Meta-Document for each class, the system computes the term weighting of all terms of the dataset of this class then the average of all terms weighting is calculated to be subsequently compared to the term weight of each term to find out whether it is higher than the average or not. Thus, to add a term to the Domain-Meta-Document, the weight of each term is compared to the average of all terms weight, if the weight of the term is greater than average then it will be added to the Domain-Meta-Document. This process recurs for each dataset generating a corresponding Domain-Meta-Document to the dataset of each class as shown in fig. 2.

The output of this algorithm is a single Domain-Meta-Document for each class that contains only terms with the highest importance based on their weights, which were selected from the classified Domain-Based training

documents. The proposed system is trained by different domain-based categories of classified documents.

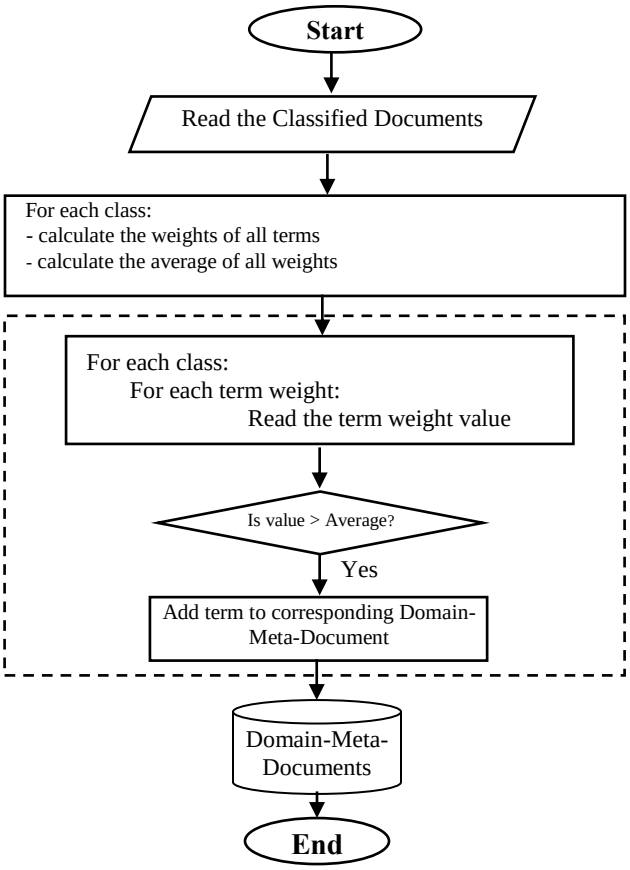

Fig. 2. The Proposed ADC Algorithm

4.5 Automated Calculation of the Threshold Value

Due to the difference in the morphological analysis of Arabic language, than that of English, the pre-processing process of Arabic data diverges of English. The obtained results of the proposed algorithms, consequently, vary based on the utilized language. Therefore, using a fixed threshold value with different languages is invalid, but there is a crucial need to find an automated and appropriate method to calculate this value according to the employed language.

Thus, Eq. (2) has been developed and proposed to achieve this essential task. The core of the proposed equation is to use the weighted average method, where the recommended threshold value for specific language is calculated based on the average rates of the selected similarities between the incoming documents and their corresponding Domain-Meta-Documents, weighted by the average values of the selected terms which were used to construct those Domain-Meta-Documents for this utilized language.

$$Thr = \frac{1}{m}\left\lfloor 1 - \varphi_j \left(\underset{j=1}{\overset{m}{Avg}} \left(\underset{i=1}{\overset{n}{Max}} \underset{j=1}{\overset{m}{Sim}} \left(Doc_i , Dom_j \right) \right) \right) \right\rfloor \quad (2)$$

Where:
Thr : The threshold value to be calculated for specific language.
m : The number of Domain-Meta-Documents.
n : Total number of documents.
Doc_i: The i^{th} document.
Dom_j: The j^{th} Domain-Meta-Document.
φ_j : Percentage of the selected terms from j^{th} training Domain to create its corresponding Domain-Meta-Document.

By using this equation we concluded that the ratio between the average of the selected terms to construct the Arabic and English Domain-Meta-Documents is roughly equal to the ratio between the calculated threshold values of those languages.

4.6 Automatic Classification Algorithm

The proposed classification algorithm depends on finding similarity for each pair of vectors, i.e. Domain-Meta-Document and input document as shown in Fig. 3.

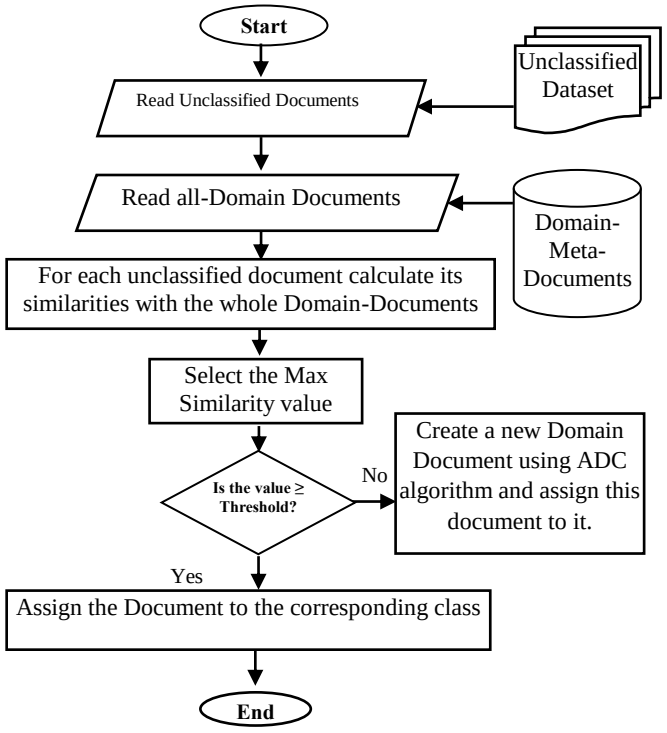

Fig. 3. The Proposed Classification Algorithm

In the beginning, the similarity between the document and all the formerly constructed Domain-Meta-Documents should be worked out, utilizing the cosine Eq. (3) which computes the cosine of the angle between two term vectors [15]. As a result, a set of similarity values for this document is computed with respect to the predefined Domain-Meta-Documents. Then the maximum similarity value of them will be selected and compared to the pre-calculated threshold similarity value in order to decide whether to assign this document to a specific class. Therefore, if the selected similarity value less than the pre-calculated threshold, the classifier will send a feedback to the ADC algorithm instigating it to create a new unlabeled Domain-Meta-Document based on the content of this document. In this case, the classifier will assign this document to unlabeled cluster automatically.

The aforementioned operation would be consecutively repeated for every incoming Arabic unclassified document with respect to all constructed Domain-Meta-Documents.

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}} \qquad (3)$$

Where:

d_j : is the vector of the weights of document j.

d_k : is the vector of the weights of Domain-Meta-Document k.

$w_{i,j}$: is the weight of the term *i* in document j.

$w_{i,k}$: is the weight of the term *i* in Domain-Meta-Document k.

4.7 Time and Space Complexity

The time and space complexity of our proposed algorithms could be discussed as follows:

Let we have N classes, each class contains M documents, each document contains L terms therefore, the classified dataset contains N*M*L terms:
$C = \{c_i: 1 \le i \le N\}$; N is the number of classes.
$D = \{d_j: 1 \le j \le M\}$; M is the number of documents in largest class.
$T = \{t_k: 1 \le k \le L\}$; L is the number of terms in largest document.

The time complexity of ADC algorithm to create a Domain-Meta-Document is divided into two steps:

Step1: *O(LM log (LM))* for each term, we need (*log LM*) to update its weight in the entire ordered list of terms.
Step2: *O(LM)* to omit all terms with weight less than the threshold.
Therefore, the total running time to create all Domain-Meta-Documents is *O(N(LM log (LM)))*.

The space complexity of ADC algorithm to create a Domain-Meta-Document is divided into two steps:

Step1: *O(L)* for each training document to store all terms and its weights in a vector space.
Step2: *O(ML)* for each training class contains M documents
Therefore, the total required space to store all of classified dataset is *O(NML)*.

In all cases no more spaces are needed, since the omitted terms will reduce the required space.

The time and space complexity of the proposed classification algorithm to compute the similarity between one document and all Domain-Meta-Documents is *O(LP)*; P is the number of dictionaries.

Finally, due to the dimensionality reduction by ADC algorithm and calculating only the similarities for P Domain-Meta-Documents and the incoming unclassified documents using the proposed classification algorithm, time and space complexities are improved.

5. Results and Evaluation

5.1 Dataset

To evaluate the proposed approach, we used two standard unstructured document sets of data. The Corpus of Contemporary Arabic (CCA) [16], one of these two sets, is composed of 12 different categories as for instance, Economic, Politics, Tourism and travel, etc., which were obtained from the online sources of Leeds University as available free of charge material.

The newsgroup 20 is the second dataset, essentially used for text classification and clustering measurement for ML techniques, contains about 20,000 articles evenly divided among 20 UseNet Discussion groups. Three groups were selected i.e. computer science, sports, and medicine in our experiment. This data set is available at http://qwone.com/~jason/20Newsgroups/.

5.2 Experimental Results

To test the Proposed ADC Algorithm which is trained by selected four classified classes, each class had a set of domain-based Arabic unstructured text files as shown in Table 1. The output of this algorithm yielded corresponding four Domain-Meta-Documents. The average number of the selected terms for all Domain-Meta-Documents was found to be about 29% in the case of Arabic language and about 23% in case of English language from all the terms for a corresponding set of the input classes so, the number of selected terms using proposed ADC algorithm in Arabic language was found to supersede the English one.

Table 2 depicts a sample of the results, which were obtained from computing the similarity degrees between Arabic or English dataset and the formerly constructed Domain-Meta-Documents. The classifier would select the underlined values (maximum values) as a key guide to classify each document to its corresponding class of Domain-Meta-Document domain.

As shown in table 1, the number of selected terms to generate the Arabic domain-meta-documents using the proposed ADC algorithm in Arabic training classes was found to be higher than the selected terms to generate English domain-meta-documents as in 29%, 23%, respectively. Therefore, when the proposed classification algorithm was applied to classify the incoming Arabic dataset, it achieved high values of similarity between the incoming documents and the generated domain-meta-documents, when compared to its counter English language case.

Table 1: The Average of the Selected Terms to Construct Each Domain-Meta-Document

Automated Domain-Meta-Documents									
Languages and Domains	**Arabic Language**					**English Language**			
	Economic	Health and Medical	Political	Tourism and Travel	**Total**	Sports	Computer	Medicine	**Total**
The Number of Terms	2287	1791	1797	1086	6961	1105	1728	3127	5960
Selected Terms	646	467	567	320	2000	198	350	801	1349
% of Selected Terms	28%	26%	32%	29%	29%	18%	20%	26%	23%

Table 2: A sample of Document to Domain-Meta-Documents Similarity Matrix values

Automated Domain-Meta-Documents								
Arabic Language					**English Language**			
Document	Economic	Health and Medical	Political	Tourism and Travel	Document	Computer	Medicine	Sports
Ec1.txt	0.58	0.42	0.50	0.53	58862.txt	0.36	0.10	0.04
Ec3.txt	0.65	0.49	0.55	0.50	60207.txt	0.46	0.16	0.10
HM1.txt	0.36	0.52	0.36	0.25	58568.txt	0.20	0.37	0.05
HM2.txt	0.36	0.51	0.34	0.29	59122.txt	0.12	0.46	0.07
Pol.txt	0.48	0.39	0.63	0.36	59123.txt	0.11	0.40	0.07
Po3l.txt	0.50	0.41	0.65	0.39	104564.txt	0.6	0.5	0.49
To1.txt	0.37	0.23	0.29	0.50	104613.txt	0.11	0.12	0.43
…..	…..	…...	…..	…..	…..	…..	…..	…..

It can be lucidly noticed that the gap between the selected similarity value and others similarity values for each document in the case of Arabic language was less than the gap value in the case of English language as shown in Fig. 4 and Fig. 5.

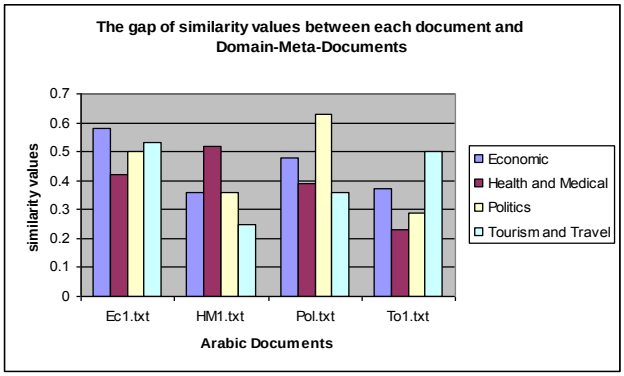

Fig. 4. The gap of yeilded similarities using Arabic Domain-Meta-Documents

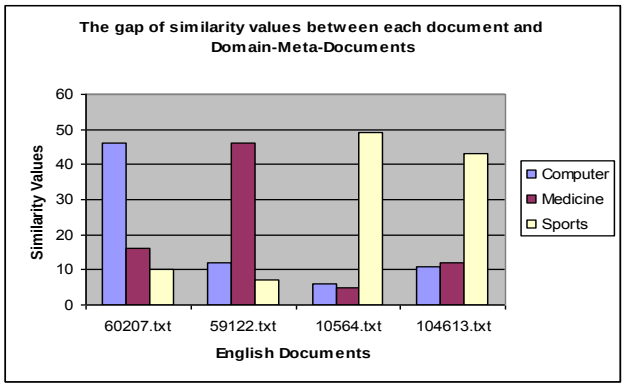

Fig. 5. The gap of yeilded similarities using English Domain-Meta-Documents

5.3 Setting the Threshold Values

According to the results obtained from the proposed Eq. (1), utilizing the similarity threshold value of 40% in case of Arabic language and of 30% in case of the English language is recommended. These recommended thresholds were found to achieve high accuracy of classification of both Arabic and English documents classification with about 95.6, 93.3%, respectively in our experiments.

5.4 Evaluation Measures

To evaluate the performance of the proposed system, standard precision, recall and F-measure are used as shown Eq. (4), Eq. (5) and Eq. (6).

$$\text{Precision} = \frac{\text{number of correctly recognized entities}}{\text{total number of recognized entities}} \quad (4)$$

$$\text{Recall} = \frac{\text{number of correctly recognized entities}}{\text{total number of correct entities}} \quad (5)$$

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (6)$$

Table 3 illustrates the results of F-measure, which were obtained from the proposed approach and they were compared to both C4.5 Algorithm in the case of Arabic language and Back Propagation Neural Network (BPNN) technique, which were presented by Fadi Thabtah in [4] and Anuradha Patra in [17], respectively. The same dataset was utilized in both cases in which the proposed system outperformed the technique to which it was compared. As for the C4.5 and BPNN the outperformance was found on F-measure.

Table 3: Comparison of Results

S.N.	Dataset	Classification Algorithm	F-measure
1	CCA	C4.5	0.899
		Proposed approach using ADC	0.956
2	Newsgroup 20	BPNN	0.92
		Proposed approach using ADC	0.933

6. Conclusions

In this paper, an improved methodology to classify Arabic/English unstructured documents was introduced. This methodology depended on two algorithms, the Automated Domain-Meta-Document Construction (ADC) and a new classification algorithm.

The obtained results revealed that the ADC algorithm had powerful ability to reduce the size of training data to about 29% in the form of automated constructed Domain-Meta-Documents. Utilizing these constructed Domain-Meta-Documents by the proposed classification algorithm based on the obtained threshold values from the developed threshold equation achieved approximately high classification accuracy (95%).

The experimental results revealed that the new proposed classification approach outperformed the compared others algorithms (C4.5 and BPNN), applied on the same dataset,

with average of F-measure (0.956, 0.899) and (0.933 and 0.92), respectively.

The logical next step in the future work is automatic labelling for the unlabeled clusters and updating the constructed Domain-Meta-Documents automatically.

References

[1] U. Kampffmeyer, M. Slominsky, and S. Pfeiffer, (2007). ECM: Enterprise Content Management. IBM Deutschland GmbH, Frankfurt. Available at: www.IBM.com/de.

[2] D. Gingell, (2006). A 15 Minute Guide to Enterprise Content Management. EMC Corp., USA.

Available at:www.EMC.com/Documentum.

[3] T. M. Mitchell. Machine Learning, New York, USA, McGraw-Hill Companies, 1997.

[4] F. Thabtah, O. Gharaibeh, and H. Abdeljaber, "Comparison of Rule Based Classification Techniques for the Arabic Textual Data", in IEEE Fourth International Symposium on Innovation in Information & Communication Technology (ISIICT), Vol., No., 2011, pp. 105-111.

[5] M. Al-diabat, "Arabic Text Categorization using Classification Rule Mining", Applied Mathematical Sciences, vol. 6, No. 81, 2012, pp. 4033-4046.

[6] M. A. H. Omer, M. S. Long, "Stemming Algorithm to Classify Arabic Documents", Symposium on Progress in Information & Communication Technology, 2009, pp. 111-115.

[7] H. Ezzat, S. Ezzat, S. El-Beltagy, and M. Ghanem, "TopicAnalyzer: A System for Unsupervised Multi-Label Arabic Topic Categorization", in IEEE International Conference on Innovations in Information Technology (IIT), Vol., No., 2012, pp. 220-225.

[8] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving Arabic Text Categorization Using Decision Trees", in IEEE First International Conference on Digital Technologies, Vol., No., 2009, pp. 110-115.

[9] S. Raheel, J. Dichy, and M. Hassoun, "The Automatic Categorization of Arabic Documents by Boosting Decision Trees", in IEEE Fifth International Conference on Signal Image Technology and Internet Based Systems, Vol., No., 2009, pp. 294-301.

[10] F. Harrag, E. El-Qawasmah, "Neural Network for Arabic Text Classification", in IEEE Second International Conference on the Applications of Digital Information and Web Technologies, Vol., No., 2009, pp. 778-783.

[11] F. Harrag, E. El-Qawasmah, A. M. S. Al-Salman, "Comparing Dimension Reduction Techniques for Arabic Text Classification Using BPNN Algorithm", in IEEE First International Conference on Integrated Intelligent Computing, Vol., No., 2010, pp. 6-11.

[12] M. F. Porter, "An Algorithm for Suffix Stripping", Program, vol. 14, No. 3, 1980, pp. 130-137.

[13] T. El-Shishtawy, and F. El-Ghannam, "An Accurate Arabic Root-Based Lemmatizer for Information Retrieval Purposes", IJCSI International Journal of Computer Science Issues, Vol. 9, No. 3, 2012, pp. 58-66.

[14] W. Zhao, Y. Wang, and D. Li, "A Dynamic Feature Selection Method Based on Combination of GA with K-Means", in 2nd International Conference on Industrial Mechatronics and Automation, vol. 2, 2010, pp. 271-274.

[15] M. Bazarganigilani, B. Fridey, and A. Syed, "Web Service Intrusion Detection Using XML Similarity Classification and WSDL Description", International Journal of u- and e-Service, Science and Technology, Vol. 4, No. 3, 2011, pp. 61-72.

[16] L. Al-Sulaiti, E.Atwell, "The Design of a Corpus of Contemporary Arabic", International Journal of Corpus Linguistics, Vol., No., 2006, pp. 1-36.

[17] A. Patra, and D. Sinph, "Neural Network Approach for Text Classification using Relevance Factor as Term Weighing Method", International Journal of Computer Applications, Vol. 68, No. 17, 2013, pp. 37-41.

Walid Mohamed Aly has acquired his Ph.D. from faculty of Engineering, Alexandria University, Egypt. He is currently working as associate professor at the College of Computing and Information Technology (CCIT), Arab Academy for Science, Technology & Maritime Transport (AASTMT). His research interests include intelligent systems, soft computing, modeling, simulation and machine learning.

Wafaa Hanna Sharaby graduated from faculty of Engineering, Ain Shams University, Egypt. He has awarded Ph.D. from Academy of Economic Studies Bucharest, ROMANIA, in the area of Management Information Systems. He is working as a lecturer at the departments of Computer Science and Information Systems, Future Academy (The Higher Institute of Specialized Technological Studies (HISTS)), Egypt. He has more than 25 years of programming, system analysis and design experience.

Hany Atef Kelleny is a student master at the College of Computing and Information Technology (CCIT), Arab Academy for Science, Technology & Maritime Transport (AASTMT). His research interests include soft computing and machine learning.