Paper: ASAT-17-121-CE

17th International Conference on
AEROSPACE SCIENCES & AVIATION TECHNOLOGY,
ASAT - 17 – April 11 - 13, 2017, E-Mail: asat@mtc.edu.eg
Military Technical College, Kobry Elkobbah, Cairo, Egypt
Tel: +(202) 24025292 – 24036138, Fax: +(202) 22621908

Towards Ontology-Based web text Document Classification

Mohamed K. Elhadad[*], Khaled M. Badran[†], and Gouda I. Salama[‡]}[§]

Abstract: The data on the web is generally stored in structured, semi-structured and un-structured formats; from the survey the most of the information of an organization is stored in unstructured textual form .so, the task of categorizing this huge number of unstructured web text documents has become one of the most important tasks when dealing with web. Categorization, Classification, of web text documents aims in assigning one or more class labels, Categories, to the un-labeled ones; the assignment process depends mainly on the contents of the document itself with the help of using one or more of machine learning techniques. Different learning algorithms have been applied on the content of text documents for the classification process. In this paper experiments uses a subset of Reuters-21578 dataset to highlight the leakage and limitations of traditional techniques for feature generation and dimensionality reduction, showing the results of classification accuracy, and F-measure when applying different classification algorithms.

Keywords: Feature Extraction, natural language processing, web text documents classification, Vector Space Mode, KNN, Principle component analysis, dimensionality reduction, term frequency inverse document frequency.

1. Introduction

Due to the increasing availability of web text documents, approximately 80% of the information of an organization is stored in unstructured textual form, and the rapid growth of the World Wide Web makes the task of automatic classification of text documents to become an interesting area for research as it is considered to be the key method for handling, managing, and organizing text data [1].

For classification process, documents are to be represented by a set of words to fully show its meaning. generally, they are processed and transformed from the full text version to a document vector, the Vector Space Model (VSM), in which each document is represented as a vector, Bag of Words (BoW)[2][3][4], which makes the handling them much easier and to reduce their complexity.

This transformation maps each document into a compact form of its content. The main problem with text documents classification is not only the extremely high dimensionality of text data, so the number of potential features often exceeds the number of training documents, but also the ignorance of the semantic information in them[5].

[*] moh.elhadad@mtc.edu.eg
[†] khaledbadran@mtc.edu.eg
[‡] gisalama@mtc.edu.eg
[§] Egyptian Armed Forces, Egypt.

This paper is organized as follows. A related work is discussed in Section 2. The main phases of the text documents classification model are introduced, in Section 3, Experimental results and performance evaluations are presented in Section 4. Finally, conclusions are given in Section 5.

2. Related Works

In this section, we briefly discuss and review some background research including the text document classification task applied to document datasets, some previous attempts for reducing the dimensionality of the used feature in the document classification process.

In [1][6][4][7][8], a full review of the current trends for text documents classification, and classification algorithms are introduced. Also [9] , comparative study of classification algorithm for text based categorization been introduced showing the advantages and dis advantages of each one of Naive Bayes, K - Nearest Neighbor, and the Decision Tree classifiers. And in[10]a technique of Building a K-Nearest Neighbor Classifier for Text Categorizationis introduced, while in [11] an improved KNN Classifier is introduced. Also in [12][13][14][15][16], dimensionality reduction techniques for enhancing automatic text categorization, and a survey of different approaches for extraction and reduction process is proposed. And [17]introduces the PCA as an efficient technique for reducing the dimensionality of Big data.

This paper proposes an approach of using PCA for reducing the dimensionality of the feature vector used for text documents classification process using traditional weighting techniques , term frequency inverse document frequency (TFIDF)[18][19],to give weights for the feature vector elements. The performance of the classification result has evaluated with the use of the F-Measure, and the Classification accuracy when using different classifiers available in the WEKA [20] datamining tool.

3. Web Text Documents Classification Model

The overall classification model passes through two stages; the Learning Stage, as explained in (3.1.), and the Classification Stage, as explained in (3.2.). The functional block diagram of the text documents classification model is depicted in Fig. 1.

3.1. The learning Stage

The first issue that needs to be addressed in text document classification is how to represent texts retaining as much information as needed without any losses. The Most commonly used for representing text documents for mining tasks is the Vector Space Model (VSM) ,in which each document is represented as a vector ,Bag of Words (BoW)[2][7]

In our approach, we use a method in which we apply aims to build the feature vector for mining tasks on both the training and the testing text documents; this is the Pre-Processing phase. The Pre-Processing phase aims in preprocess the input documents; extracting the BoW that represents these documents by performing set of steps such as Natural Language Processing (NLP) Parser, to detect words from the document phrases, removing the stopping words, and clean the data from noisy words that contain symbols and non-English characters, finally perform the stemming process to replace each extracted word by its morphological root as mentioned in the related works.

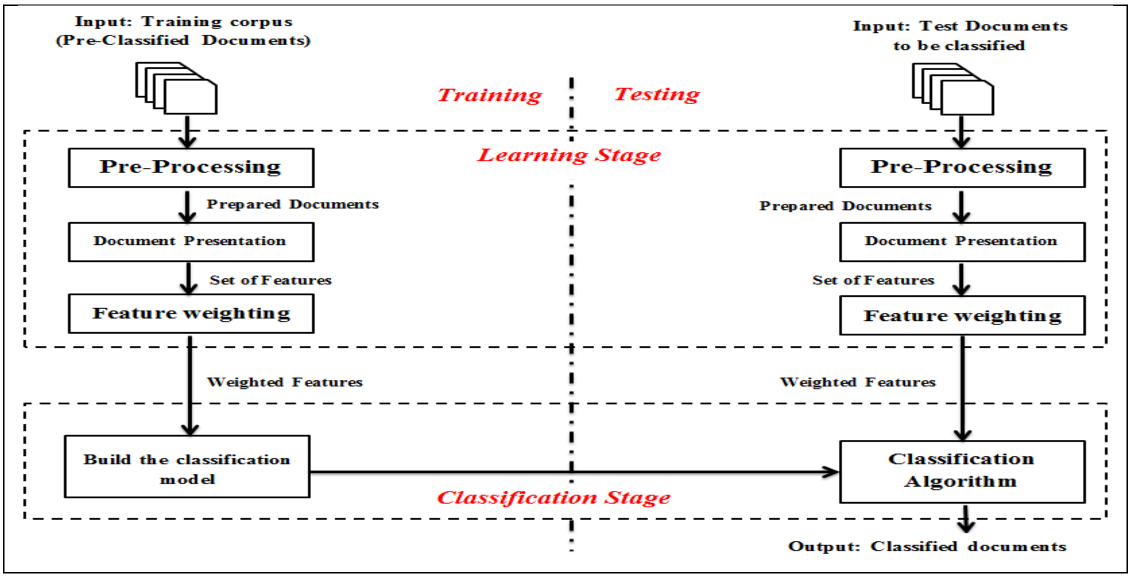

Fig. 1.Block diagram of the text documents classification model

After that, the extracted set of features is weighted to indicate the importance of each feature by using the TFIDF as a weighting technique [3]. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

$$\mathrm{TFIDF}(t) = \mathrm{TF}(\mathrm{T}) \times \mathrm{IDF}(t) \qquad (1)$$

where TF measures how frequently a term occurs in a document and the IDF measures how important a term is, as follows [3]:

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}} \qquad (2)$$

$$\mathrm{IDF}(t) = \frac{\text{log(Total number of documents)}}{\text{Number of documents with term t in it}} \qquad (3)$$

This means that larger weights are assigned to terms that appear relatively rarely throughout the corpus, but very frequently in individual documents.

3.2. The Classification Stage

For the text document classification purpose, the open source Weka data mining software suite was utilized. Weka is based on Java programming language and it has a collection of machine learning algorithms for data mining tasks. The users have the capability to incorporate the Weka code base into their own Java code and call the Weka libraries for the classification purposes. Weka is very well known for its rich set of libraries that can be utilized for machine learning purposes.

The classification is final stage to get the decision. In this research, five different classification schemes were carried out in an attempt to find the most suitable classifier. These classifiers, included in WEKA data mining tool, are: Naive-Bayes, J48, JRip, SVM, and KNN using different distance measures such as Euclidean distance, Manhattan distance, and Minkowski distance.

4. Experimental Results and Discussion

In this section, we discuss our experimental setup and the results for evaluating the performance of our text documents classification approach using. The experiments are divided into two parts. The first part works on reuters-21578 dataset [21] and evaluates the classification system by applying the classifiers as mentioned in 3.2., applying TFIDS as one of the traditional feature weighting techniques without any reduction of the extracted feature vector, while in the second part we use PCA as a feature reduction technique in the document presentation step.

4.1. Dataset Used

The Reuters-21578 dataset has been used in many text categorization experiments; the data was collected by the Carnegie group from the Reuters newswires in 1987. It consists of 21578 collections of new stories classified into topics. However, not all documents have a topic, there exist some of the documents that have more than one topic, and not all documents have a text in the news body.

So, we used only documents that have only one topic and have text in its body, ignoring those which have no text in the body and associated with more than one topic. The dataset has classes of different sizes; some classes have large size such as earn class that has 3734 documents belonging to it, while other classes have size less than 5 documents such as rice. So, in our experiments, we used a reduced, unbiased subset from this corpus as our dataset for training and testing each group contains between 30 and 100 documents as indicated in Table 1.

Table 1: Number for training and testing documents used

Category	#Training	#Test	#Total
gold	70	20	90
money-supply	70	17	87
gnp	49	14	63
cpi	45	15	60
cocoa	41	12	53
alum	29	16	45
grain	38	7	45
copper	31	13	44
jobs	32	10	42
reserves	30	8	38
rubber	29	9	38
iron-steel	26	11	37
ipi	27	9	36
nat-gas	22	11	33
veg-oil	19	11	30
#Total	558	183	741

4.2. Evaluation Criteria

All documents for training and testing Passes through the stages in section 3, Experimental results reported in this section are based on F1 measures and Classification accuracy.

The F1 measure is the harmonic mean of precision and recall as follows [22]:

$$F_1(\text{recall. precision}) = \frac{2\times\text{recall}\times\text{precision}}{\text{recall}+\text{precision}} \tag{5}$$

In the above formula, precision and recall [22]are two standard measures widely used in text categorization literature to evaluate the algorithm's effectiveness on a given category where

$$\text{precision} = \frac{\text{true positive}}{\text{true positive}+\text{false positive}} \tag{6}$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive}+\text{false negative}} \tag{7}$$

$$\text{Accuracy} = \frac{\#\text{ of Correctly classifieddocuments}}{\text{total number of documents}} \tag{8}$$

4.3. Experimental Results

To select the most suitable classifier from the implemented classifiers : Naive-Bayes, J48, JRip, SVM, and KNN using different distance measures such as Euclidean distance, Manhattan distance, and Murkowski distance; an experiment was implemented on the reuters-21578 dataset before and after applying PCA feature reduction technique. Table2, shows a comparison between the Feature Vector size before and after applying the PCA for dimensionality reduction. The PCA reduces numbers of features used across the dataset from 4006 to 512 features. It means that the dataset compressed with a **Compression Ratio** = 4006 / 512 = **7.824**

To study the effect of the reduced dataset feature vector by deriving a new feature vector from the available feature vector of reuters-21578 dataset, the accuracy and f-measure of the five classifiers are evaluated.

Table 2. Reduction ratio when applying PCA for Dimensionality reduction Versus the Original vector size

Feature vector	Vector size	Reduction %
Original	**4006**	**0.0 %**
Reduce based on PCA	**512**	**87.22%**

Figure 2(a,b) shows the comparison of accuracy and F-measure respectively for the five classifiers using the data set before and after applying PCA feature reduction technique. The test validation method based on dividing dataset into 70% as training set and 30% as testing. It could be noticed that, the accuracy of SVM classifier before applying the PCA for feature reduction (86.036%) is much better than the other classifiers. Also, it could be seen that, the accuracy of J48 classifier after applying the PCA for feature reduction (49.5495%) is much better than the other classifiers.

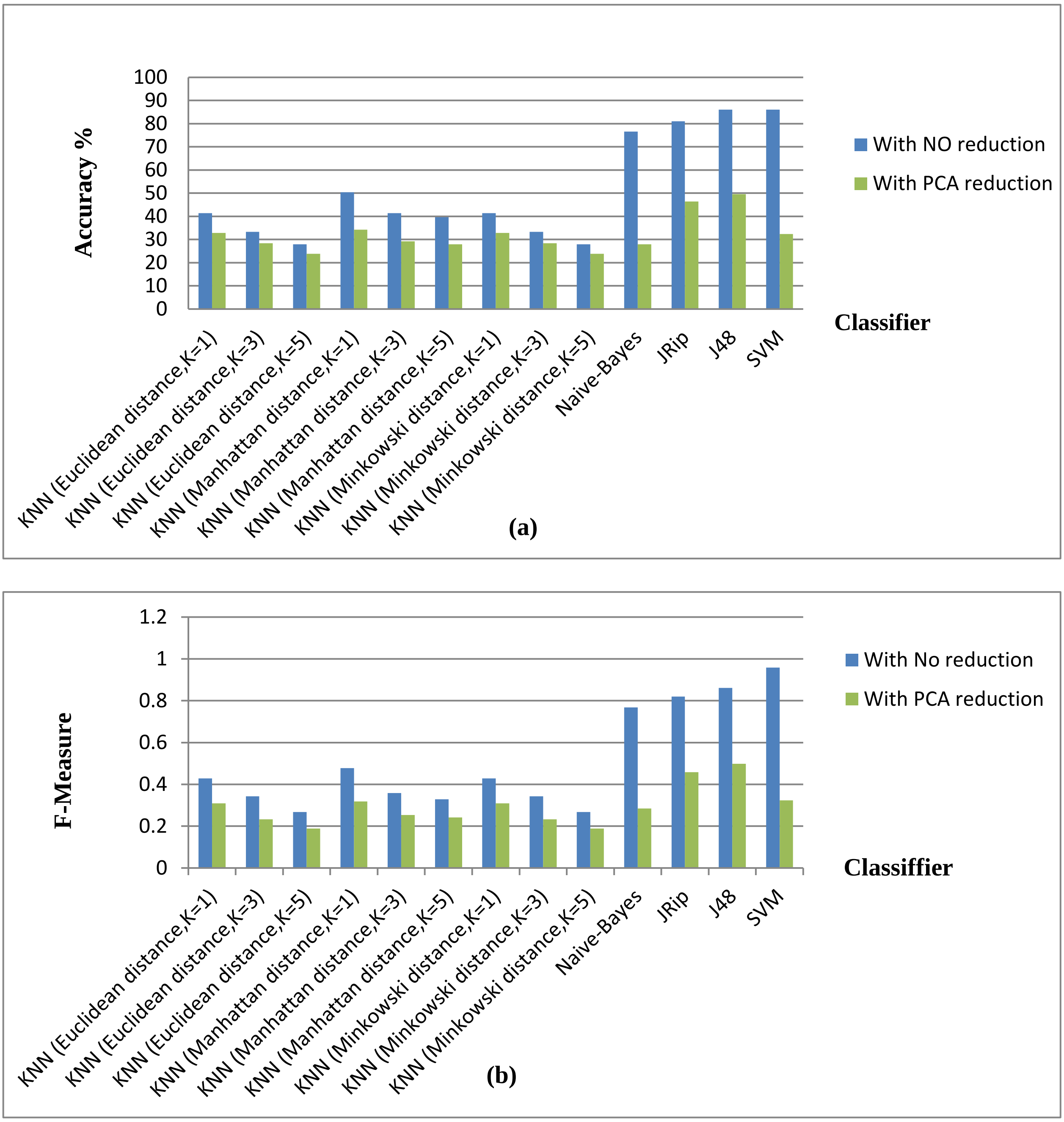

Fig.2. Evaluation measures using five different classifiers
(Naive-Bayes, JRip, J48, SVM, and KNN).
a) Classification Accuracy b) F-Measure

PCA ***is*** a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [17].

In summary, Applying the PCA technique for feature reduction will reduce the feature vector size and decrease the classification accuracy and F-measure as shown in Fig 2(a,b). The classification accuracy and F-measure of the implemented five classifiers before applying PCA outperforms the results after applying PCA because the features of the data set documents are highly correlated. So, the number of principal components is less than the number of original features that represent the document.

5. Conclusion

In this paper, we performed an experimental evaluation on Reuters-21578 dataset comparing result from using the original feature vector ,without any reduction, against the reduced one ,using the PCA as one of the corresponding classical dimensionality reduction methods, as an input for different classifiers. The experiments shows that traditional techniques used for reducing the dimensionality of the used feature vector for web text documents classification such as PCA gives high reduction ratios but it affects badly the classification results.

In future work, we recommend extending this work by utilizing ontologies such as the WordNet Ontology to perform the reduction of the feature vector instead of using the traditional reduction technique to enhance the results of the different classifiers and overcome the shortcomings of the traditional techniques.

6. References

[1] B. B. B. K. K. Aurangzeb Khan, "An Overview of E-Documents Classification," in *International Conference on Machine Learning and Computing*, Singapore, 2011.

[2] D. S. G. S. M. B S Harish, "Representation and Classification of Text Documents: A Brief Review," *IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition",* 2010.

[3] P. R. ,. a. H. S. Christopher D. Manning, Introduction to Information Retrieval, Cambridge : Cambridge University Press, 2008.

[4] a. D. V. P. Sayali Rasane, "Handling Various Issues In Text Classification : A Review," *International Journal on EmergingTrends in Technology (IJETT),* vol. 3, no. 1, pp. 4076-4082, 2016.

[5] a. T. G. Kerem Çelik, "A Comprehensive Analysis of using Semantic Information in Text Categorization," in *The IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA 2013)*, Albena, 2013.

[6] R. M. A. J. Rajni Jindal, "Techniques for text classification: Literature review and current trends," *Webology,* vol. 12, no. 2, p. Article 139., 2015.

[7] S. a. C. C.Uma, "A Survey Paper on Text Mining Techniques," *International Journal of Engineering Trends and Technology (IJETT),* vol. 40, no. 4, pp. 225-229, 2016.

[8] L. P. a. N. M. N. Venkata Sailaja, "Survey of Text Mining Techniques, Challenges and their Applications (IJCA)," *International Journal of Computer Applications,* vol. 146, no. 11, pp. 30-35, 2016.

[9] G. P. D. U. K. J. Omkar Ardhapure, "Comparative Study Of Classification Algorithm For Text Based Categorization," *International Journal of Research in Engineering and Technology (IJRET),* vol. 5, no. 2, pp. 217-220, 2016.

[10] K. R. A.Kousar Nikhath, "Building a K-Nearest Neighbor Classifier for Text Categorization," *International Journal of Computer Science and Information Technologies (IJCSIT),* vol. 7, no. 1, pp. 254-256, 2016.

[11] a. R. R. Shaifali Gupta, "Improvement in KNN Classifier (imp-KNN) for Text Categorization," *International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE),* vol. 6, no. 6, pp. 276-280, 2016.

[12] J. L. ,. Q. Z. a. Y. W. Haozhe Xie, "Comparison among dimensionality reduction techniques based on Random Projection for cancer classification," *Computational Biology and Chemistr,* vol. 65, p. 65–172, 2016.

[13] a. S. K. R. Masoumeh Zareapoor, "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection," *international journal Information Engineering and Electronic Business,* vol. 2, pp. 60-65, 2015.

[14] M. W. Mwadulo, "A Review on Feature Selection Methods For Classification Tasks," *International Journal of Computer Applications Technology and Research,* vol. 5, no. 6, pp. 395-402, 2016.

[15] a. M. M. Pradnya Kumbhar, "A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification," *International Journal of Science and Research (IJSR)* , vol. 5, no. 5, pp. 1267-1275, 2016.

[16] D. A. Said, "Dimensionality Reduction Techniques For Enhancing Automatic Text Categorization," *Master Thesis,Cairo University,* 2007.

[17] a. B. Y. Tonglin Zhang, "Big Data Dimension Reduction using PCA," in *IEEE International Conference on Smart Cloud,* New York, 2016.

[18] a. J. Y. M. Liu, "An improvement of TFIDF weighting in text categorization," Singapore, 2012.

[19] H. C. L. R. W. P. W. K. F. &. K. K. L. Wu, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS),* vol. 26, no. 3, p. 13, 2008.

[20] M. L. Group, "Downloading and installing Weka," The University of Waikato, [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/downloading.html. [Accessed 10 January 2017].

[21] "The Reuters dataset is available to be downloaded in sgml format from," [Online]. Available: http://www.daviddlewis.com/ressources/testcollections/reuters21578/. [Accessed 12 January 2017].

[22] a. G. L. Marina Sokolova, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management,* vol. 45, no. 4, p. 427–437, 2009.