Using Fuzzy LR Numbers in Bayesian Text Classifier for Classifying Persian Text Documents

Parisa Pourhassan
M.A. of Information Technology Management, Islamic Azad University, E-Campus,
No.15, Ali St., Ghaemshahr, Mazandaran, Iran
+98 911 1298878
Parisa_pourhassan@yahoo.com

Alireza Pourebrahimi
Assistant Prof. of Islamic Azad University, Karaj Branch, Iran
No.313, Rajaee shahr St., Karaj, Iran
+98 26 34418143
poorebrahimi@gmail.com

Mohammad Ali Afshar Kazemi
Associate Prof. of Islamic Azad University, Central Branch
No.166, Zafar St., Tehran, Iran
+98 912 3336731
drafshar@iauec.com

ABSTRACT

Text Classification is an important research field in information retrieval and text mining. The main task in text classification is to assign text documents in predefined categories based on documents' contents and labeled-training samples. Since word detection is a difficult and time consuming task in Persian language, Bayesian text classifier is an appropriate approach to deal with different word formats and new words. Also, fuzzy theory may be used to manage uncertainty in imprecise Persian sentences.

In this paper, we utilize L-R type fuzzy numbers in Bayesian text classifier to classify textual Persian documents (Fuzzy Bayesian text classifier). The obtained results on simulated imprecise textual Persian documents show improvements in both recall and precision parameters by using Fuzzy Bayesian text classification approach over Naïve Bayesian text classifier.

Keywords

Text Classification, Fuzzy L-R Numbers, Bayesian Classification.

1. INTRODUCTION

Rapid development of the internet has made a large number of computer-readable text information available. In order to effectively manage and utilize the large amount of documents, many researches are working on automatic classification and categorization of documents.

Artificial intelligence is an area that utilizes computational techniques and methodologies to perform complex tasks with great performance and high accuracy [1]. During recent years, the majority of researches have investigated on text classification through supervised machine learning techniques [2,3,4].

Text Mining (TM) is a multi-disciplinary area in machine learning that needs general knowledge about computing, statistics, probability and linguistics [5].

Text Classification (TC) is an important research field in information retrieval and text mining. Main task in text classification is assigning textual documents to predefined categorizes based on documents' contents and labeled training samples [6]. It is used in Natural Language Processing (NLP) to process textural data by finding out their grammatical syntax and semantics and representing them in a fully structured form [1,7].

Text Mining and text classification are applied in various applications such as: web news classification [8], language identification [9], spam filtering [10,11,12], medical document categorization [13,14], and so on.

On the other hand, a statistical approach based on feature extraction from labeled- training samples is a popular approach to generate a knowledge base with minimum cost [15]. Besides, in a test administration phase, statistical techniques like Naïve Bayesian classifier works with liner complexity order. Hence, statistical approaches are employed in variety of researches to solve the text classification problem [16].

There are many imprecise sentences in terms of meaning for an English native speaker in Persian language; for example, when somebody says "gololeh khordeh ast.", it means "He is shot" not "He has eaten a bullet". In this paper, we apply fuzzy theory with naïve Bayesian classifier to overcome ambiguity problem in Persian sentences within text classification.

As word derivation is a challenging process in Persian language and converting verbs to different tenses is irregular,

the researchers utilize naïve Bayesian classifier that can adapt to this condition.

2. RELATED WORKS

Dharmadhikari and his colleagues presented a review of various text classification approaches under machine learning paradigm [17].

Puri suggests a Fuzzy similarity based on the concept of Mining Model (FSCMM) to classify text documents into predefined categories. She performs text classification by srutinizing the sentences, documents and integrated corpora levels along with feature reduction and ambiguity removal on each level to achieve a high system performance [2].

Alsaleem has applied Naïve Bayesian (NB) method and Support Vector Machine (SVM) algorithm to render classification of different Arabic data set. Experimental results indicate SVM algorithm surpasses the NB in terms of overall criteria [3].

Zhou and his coworkers have suggested an improved KNN text classification algorithm based on clustering. They compress the given training set and delete the samples near the border leading to the elimination of the multi-peak effect of the training sample set. Thun, training sample sets of each category are clustered by k-means clustering algorithm and all cluster centers are taken as the new training samples. Also, A weight value indicates the importance of each training sample compatible with the number of samples in the cluster comprising the cluster center. Finally, modified samples are used to accomplish KNN text classification. Simulations results of this approach show that the proposed algorithm can effectively reduce the actual number of training samples and also the calculation complexity [4].

Krishnalal and his colleagues have proposed an intelligent system for online news classification based on Hidden Markov Model (HMM) and Support Vector Machine (SVM). An intelligent system is designed to extract the keywords from online newspaper contents based on HMM feature extraction and to classify it according to the predefined categories using SVM. Experimental results are evaluated as satisfactory compared to other text classification methods [8].

Text classification approaches are applied to spam recognition in [10,11,12]. Zhang and her colleagues have evaluated various statistical spam recognition approaches such as naïve Bayesian, Support Vector Machine, Maximum Entropy Model [10]. Sabri and her co-workers have suggested a spam recognition approach based on modified Artificial Neural Networks (ANN) [11]. They called their approach Continuous Learning Approach Artificial Neural Networks (CLA_ANN). Subramanian and his colleagues have summarized the most common techniques used for spam recognition by analyzing e-mail contents. They used machine learning algorithms such as Naïve Bayesian, Support Vector Machine and Neural Networks that have been adopted to detect and control spam [12].

Zurini and her co-workers suggest a personal approach based on spatial dimension model in the process of classification. They also recommended an approach to spam recognition [18]. Also Zhu and her colleagues suggest an approach to categorize the enriched format text using component similarity [19]. They have obtained feature structure distribution weight in their research. They also take text formats in feature weighting into account. Finally, they categorized texts in terms of similarity in document components.

3. PROPOSED METHOD

3.1.Background

A Bayesian classifier is simply a Bayesian network applied to a classification task. It contains a node C to representing the class variable and a node X_i for each feature. Given a specific instance x (an assignment if values $x_1, x_2, \dots, x_n$ to the feature variables), the Bayesian network allows us to compute the probability $P(C = c_k | X = x)$ for each predefined class c_k. Equation (1) shows how this probability is calculated based on Bayes theorem.

$$P(C = c_k | X = x) = \frac{P(X = x | C = c_k) P(C = c_k)}{P(X = x)} \quad (1)$$

The critical quantity in Equation (1) is $P(X = x | C = c_k) P(C = c_k)$, which is often impractical to compute without imposing independent assumption. The old and most restrictive form of such assumption is embodied in the Naïve Bayesian classifier assuming each X_i is conditionally independent of every other feature, given the class variable C formally, these yields

$$P(X = x | C = c_k) = \prod_i P(X_i = x_i | C = c_k) \quad (2)$$

Fuzzy logic [19][20] is a tool to deal with uncertain, imprecise, or qualitative decision-making problems. Unlike Boolean logic, where an element x either belongs or does not belong to a set A, in fuzzy logic the membership of x in A has a degree value in a continuous interval between 0 and 1.

In the other word; if X be a nonempty set. A fuzzy set A in X is characterized by its membership function.

$$\mu_A : X \to [0,1] \quad (3)$$

And $\mu_A(x)$ is interpreted as the degree of membership of elements x in fuzzy set A for each $x \in X$ [19] [20].

It is clear that A is completely determined by the set of Tuples

$$A = \{(u, \mu_A(u)) | u \in X\} \quad (4)$$

If $X = \{x_1, x_2, \dots, x_n\}$ is a finite set and A is a fuzzy set in X then we often use the notation in equation (5) to describe A [19].

$$A = \sum_{i=1}^{n} \mu_i / x_i \quad (5)$$

Where the term $\mu_i / x_i, i = 1, \dots, n$ signifies that μ_i is the grade of membership of x_i in A and the sigma sign represents the union. Also for an infinite set X, to describe fuzzy set A we can use notation in equation (6) [19].

$$A = \int \mu_i / x_i \quad (6)$$

A fuzzy subset A of a classical set X is called normal if there exists and $x \in X$ such that $A(x) = 1$, otherwise A is subnormal. A fuzzy set A of X is called convex if $\alpha - cut(A)$ is a convex subset of $X \, \forall \alpha \in [0,1]$ [20].

A fuzzy number A is a fuzzy set of the real line with a normal, convex and continuous membership function of bounded support [20].
A fuzzy number M is said to be an L-R fuzzy number if

$$\mu_M(x) = \begin{cases} L\left(\frac{m-x}{\alpha}\right) \; if \; x \le m, \alpha > 0 \\ R\left(\frac{x-m}{\beta}\right) \; if \; x \ge m, \beta > 0 \end{cases} \quad (7)$$

Where m is the mean value of M, α and β are called left and right spreads. Also in equation (7) $L(x)$ is the left and $R(x)$ is the right references that have to obey following conditions:

- R(x)=R(-x), L(x)=L(-x)
- L(0)=1 , R(0)=1
- L and R are non-increasing on [0, +∞]

Symbolically we write $M = (m, \alpha, \beta)_{LR}$ to show an L-R type fuzzy number [19].

3.2 Fuzzy Bayesian Text Classifier

Our proposed approach for classifying text documents has two major phases, Text Learning Phase and Text Classification Phase. Labeled text documents are used for knowledge extraction in text learning phase and in text classification phase, fuzzy Bayesian approach are used to classifying new text documents in predefined classes based on extracted knowledge in text learning phase. The proposed Fuzzy Bayesian Text Classifier Model is presented in Figure 1.

Considering a set of n labeled text documents,

$$TD_C = \{\langle TD_1, C_i\rangle, \langle TD_2, C_j\rangle, \dots, \langle TD_n, C_m\rangle\} \quad (8)$$

Where $TD_1, TD_2, \dots, TD_n$ are the individual and independent text documents with specified class C_k, Text Document Training Processor (TDTP) in learning phase performs followinfg tasks on each text document:

- Text Preprocessing: TDTP processes the TD_i to prepare it to knowledge extraction in next steps. This process includes removing stop words and other extra terms except the nouns, proper nouns and numerals. To remove the invalid and extra words, the Pseudo Thesaurus is used. Because word derivation is very difficult task in Persian language, no derivation is performed in this step.
- Feature Extraction: Each text document $\langle TD_i, C_k\rangle$ is composed of features that can be used to training classifier. In fact a set of y features $F_{ik} = \{f_{1ik}, f_{2ik}, \dots, f_{yik}\}$ observed in a specific class will be extracted in this step.
- Feature Frequency Extraction: TDTP traces the extracted feature set to find the frequency of each feature. Hence, in this step a number that shows the feature frequency will be attached to each feature in F_{ik} ($F_{ik} = \{\langle f_{1ik}, r_{1ik}\rangle, \dots, \langle f_{yik}, r_{1ik}\rangle\}$).

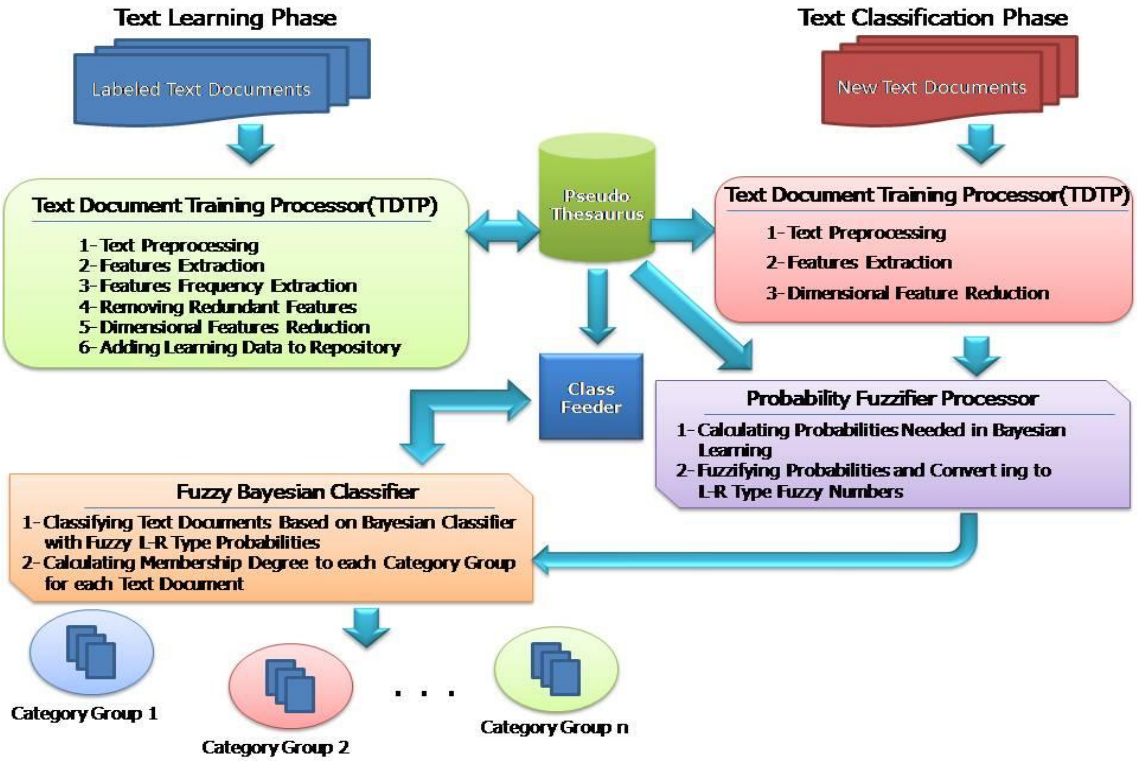

Figure 1. Fuzzy Bayesian Text Classification Model

- Removing Redundant TDTP removes redundant feature and just keeps an instance of each different feature.
- Dimensional Feature Reduction: Features with low frequency will be omitted to reduce feature set dimension.
- Adding Learning Data to Repository: Extracted features with relevant frequencies will be added to repository. If a feature existed in database previously, its frequency would be updated; otherwise, it would be inserted in database with its frequency extracted for C_k. Also, count of observation of C_k would be incremented in database.

Also, given a set of new text document

$$TD = \{TD_1, TD_2, \dots, TD_m\} \quad (9)$$

Where $TD_1, TD_2, \dots, TD_n$ are the individual and independent text documents, TDTP in classification phase performs text preprocessing, feature extraction and dimensional feature reduction tasks on each text document TD_i described previously in leaning phase results in a feature vector F_i ($F_i = \{F_{1i}, F_{2i}, \dots, F_{yi}\}$). This feature vector will be utilized with Probability Fuzzifier Processor to calculation and fuzzifying probabilities needed for Bayesian Classifier.

Probability Fuzzifier Processor receives feature extracted from TD_i ($F_i = \{f_{1i}, f_{2i}, \dots, f_{yi}\}$) and searches repository for frequency of each feature f_{li} in all predefined classes C_k and frequency of observation of type C_k text documents in learning phase. The Probability Fuzzifier Processor converts probabilities to fuzzy L-R type numbers. To achieve this purpose, Probability Fuzzifier Processor has to determine *m*, α, β, *L(x)* and *R(x)*. Equation (10) presents *L(x)* and *R(x)* that are used in this paper.

$$L(x) = R(x) = max(0, 1 - |x|) \quad (10)$$

α and β are initialized as presented in equation (11). These variables are transient and will be changed continuously in calculations.

$$\alpha = min(0.5, m)$$
$$\beta = \min(0.5, 1 - m) \quad (11)$$

Finally, Probability Fuzzifier Processor has to compute *m* for probabilities needed with Bayesian classifier. If $\check{P}(f_{li}|C_k)$ be probability of observing f_{li} in class C_k based on learning text documents, we can calculate *m* for L-R type fuzzy number $\check{P}(f_{li}|C_k)$ as

$$m = {}^{r_{kli}}\Big/_{\sum_{j=1}^{h} r_{klj}} \quad (12)$$

Where r_{kli} is observation frequency of f_{li} in class C_k and *h* is number of individual features observed in class C_k based on learning text documents. Also if $\check{P}(C_K)$ be probability of observing a text document labeled C_k based on learning text documents, m for L-R type fuzzy number $\check{P}(C_K)$ could be computed as

$$m = {}^{r_k}\Big/_{\sum_{j=1}^{d} r_j} \quad (13)$$

Where r_k is observation frequency of text documents labeled with C_k based on learning text documents and *d* is number of classes.

Now Fuzzy Bayesian Classifier can estimate maximum a posterior probability and find the most probable class for each text document TD_i.

$$C_{MAP} = arc_{c \in C} max \tilde{P}(c|\, TD_i)$$
$$= arc_{c_k \in C} max \left(\tilde{P}(c_k) \prod_{j=1}^{y} \tilde{P}(f_{ji}|c_k) \right) \quad (14)$$

Where f_{ji} is a feature extracted from TD_i and *y* is number of feature extracted from TD_i . Finally Fuzzy Bayesian Classifier will compute membership degree of each text document TD_i to specific class C_k.

$$\mu_{C_k TD_i} = \frac{\tilde{P}(c_k) \prod_{j=1}^{y} \tilde{P}(f_{ji}|c_k)}{\sum_{l=1}^{d} \tilde{P}(c_l) \prod_{j=1}^{y} \tilde{P}(f_{ji}|c_l)} \quad (15)$$

4. EXPERIMENTAL RESULTS

In this section the obtained results ,after simulating the proposed approach, are presented in diagrams and tables. Proper data is collected from simulations to evaluate the respective performance and dependability of the parameters during simulations.

Since there is no standard Persian test set to evaluate proposed approach, the researchers have chosen a text set drawn from Persian online newsletters to evaluate effectiveness of new approach, including five major categories: sports, finance, politics, social and science

The first evaluated parameter in this paper is Response Time that is a performance -oriented parameter.

Figure 2 presents response time vs. number of classes and Figure 3 presents response time vs. number of features for both naïve Bayesian and Fuzzy Bayesian text classifier.

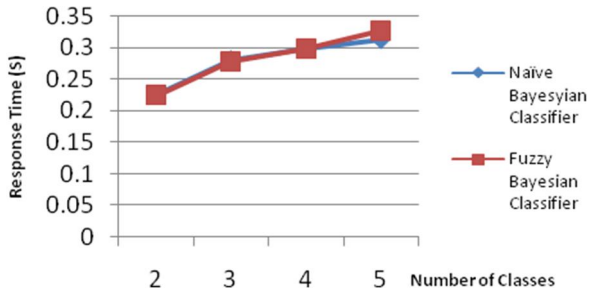

Figure 2. Response Time vs. Number of classes

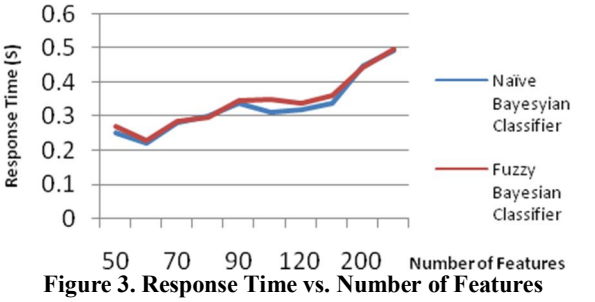

Figure 3. Response Time vs. Number of Features

The results yielded from simulations for response time shows that computation complexity for both Fuzzy Bayesian text classifier and naïve Bayesian classifier is the same and due to the complexity of basic operations in L-R type fuzzy numbers little differences are justifiable.

Other evaluated parameter is precision. Precision is defined as the percentage of text documents classified in correct class. Figure (4) presents precision evaluated from simple text documents for both Fuzzy Bayesian text classifier and naive Bayesian text classifier.

We use simple text documents for documents with no uncertain or imprecise sentences. As presented in figure (4), Fuzzy Bayesian text classifier has a little better precision compared with naïve Bayesian text classifier. Finally precision for imprecise text documents is evaluated during simulations. Figure (5) presents simulation results for evaluating precision in imprecise text documents.

As presented in figure (5), there is a significant difference in precision evaluated for Fuzzy Bayesian text classifier and precision evaluated for naïve Bayesian classifier during imprecise text documents classification. The former reaches over 98% , the latter is about 81%.

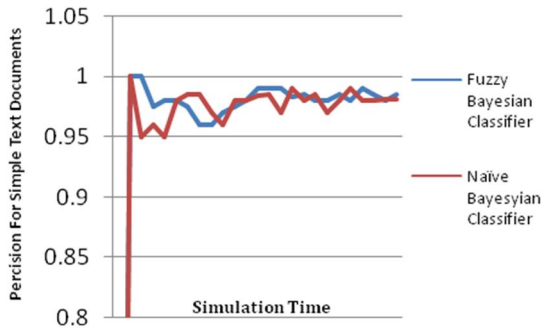

Figure 4. Precision vs. Simulation Time during Simple Text Documents classification for Both Fuzzy Bayesian and naive Bayesian Text Classifier

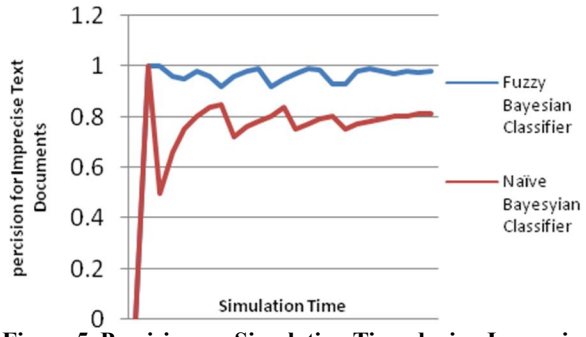

Figure 5. Precision vs. Simulation Time during Imprecise Text Documents Classification for Both Fuzzy Bayesian and naive Bayesian Text Classifier

Precision and recall parameters evaluated during simple text documents classification are summarized in Table 1. Recall is defined as the proportion of the number of correctly classified text documents to the number of text documents originally classified in a specific class.

Table 1. Precision and Recall Parameters Evaluated for Fuzzy Text Classifier and Naive Bayesian Text Classifier during Simple Text Documents Classification

Classification Approach	Precision	Recall
Fuzzy Bayesian	98.5%	97.3%
Naïve Bayesian	98.1%	97.05%

Also, Table 2 presents precision and recall parameters evaluated during imprecise text documents classification for Fuzzy Bayesian text classifier and naïve Bayesian text classifier.

Table 2. Precision and Recall Parameters Evaluated for Fuzzy Text Classifier and Naive Bayesian Text Classifier during Simple Text Documents Classification

Classification Approach	Precision	Recall
Fuzzy Bayesian	98%	98.6%
Naïve Bayesian	81%	80.34%

5. CONCLUSION

Owing to significant imprecision in Persian language sentences, Fuzzy Bayesian text classification approach can contribute to overcome the uncertainty.

Simulation results show improvement in both recall and precision parameters by using Fuzzy Bayesian text classification approach compared to naïve Bayesian text classifier during imprecise text documents classification in Persian language.

Moreover, simulation results indicate response time for naïve Bayesian classifier is a little better than Fuzzy Bayesian text classifier.

REFERENCES

[1] Padhy, N. P.(2009). Artificial Intelligence and Intelligent Systems, 5th edition, Oxford University Press.

[2] Puri, S.(2011). A Fuzzy Similarity Based Concept Mining Model for Text Classification, *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol.2, No.11, pp. 115 - 121.

[3] Alsaleem,S.(2011) Automated Arabic Text Categorization Using SVM And NB, *International Arab Journal of e-Technology*, Vol. 2, No. 2, pp. 124-128.

[4] Zhou, Y., Li, Y., & Xia, S. (2009). An improved KNN text classification algorithm based on clustering. Journal of computers, 4(3), 230-237.

[5] Han, J., & Kamber, M.(2006). Data Mining: Concepts and Techniques, 2th Edition, Elsevier.

[6] Jinshu, S., Bofen ,Z., & Xin, X.(2006). Advences in Machine Learning Based Text Categorization, *Journal of Software*, Vol. 17, No. 9, pp. 1848-1859.

[7] Rich, E., Knight, K., & Nair, S.B.(2010). Atificial Intelligence, 3th Edition, Mc Graw Hill.

[8] Krishnalal, G., Rengarajan, S. B., & Srinivasagan, K.G.(2010). A New Text Mining Approach Based on HMM-SVM for Web News Classification,

International Journal of Computer Applications, Vol. 1, No. 19, pp. 98-104.

[9] Shehata, S., Karray, F.,& Kamel, M.S.(2010). An Efficient Concept-Based Mining Model for Enhancing Text Clustering, *IEEE Transactio on Knowledge And Data Engineering*, Vol. 22, No. 10, October 2010.

[10] Zhang, L., Zhu, J., & Yao, T.(2004). An Evaluation of Statistical Spam Filtering Techniques, ACM Transaction on Asian Language Information Processing, Vol. 3, No. 4, pp. 243-269. December.

[11] Sabri, A.T., Mohammads, A.H., Al-Shargabi, B .,& Hamdeh, M. A.(2010). Developing New Continuous Learning Approach for Spam Detection Using Artifical Neural Netwok (CLA_ANN), European Journal of Scientific Research, Vol. 42, No. 3, pp. 525-535.

[12] Subramanian, T., Jalab, H. A.,& Taqa, A.Y.(2010). Overview of textual anti-spam filtering techniques, *International Journal of the Physical Sciences*, Vol. 5, No. 12, pp. 1869-1882, October.

[13] Cohen, A.M.(2006). An Effective General Purpose Approach for Automated Biomedical Document Classification, AMIA Annual Symposium Proceedings, pp. 161-165.

[14] Yildiz, M.Y.,& Pratt, W.(2005). The Effect of Feature Representation on MEDLINE Document Classification, AMIA Annual Symposium Proceedings, pp.849-853

[15] Dalal, M.K.,& Zaveri, M.A.(2011). Automatic Text Classification: A Technical Review, *International Journal of Computer Applications*, Vol. 28, No. 2, pp. 37-40, August.

[16] Manning, C.D., Raghavan, P.,& Schutze, S.(2008). Introduction to information Retrieval, 1st Edition, Cambridge University Press.

[17] Dharmadhikari, S.C., Ingle, M.,& Kulkarni,P.(2011). Empirical Studies on Machine Learning Based Text Classification Algorithms, Advanced Computing: *An Intenational Journal (ACIJ),* Vol. 2, No. 6, pp. 161-169, November.

[18] Zurini, M., Sbora, C.(2011). Clustering Analysis within Text Classification Techniques, *Informatica Economica,* Vol. 15, No. 4, pp. 178-188.

[19] Zhu, F.,& Zhou, Y.(2011). Enriched Format Text Categorization Using A Component Similarity Approach, *Journal Of Software*, Vol. 6, No. 9, pp. 1713-1720, September.

[20] K. Tanaka, K.(1996). An Introduction to Fuzzy Logic for Practical Applications, 1st Edition, Springer, New York.

[21] Alavala, C.R.(2008). Fuzzy Logic and Neural Networks: Basic Concepts and Applications, 1st Edition, New Age International Pvt Ltd Publishers, December