ISSN No. 0976-5697

Volume 3, No. 1, Jan-Feb 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# Text mining of Document using Keyphrase Extraction and Artificial Neural Network Approach for Machine Learning

Shobha Sanjay Raskar*
Bharati Vidyapeeth University, BVUCOE, Department of computer Science and Engineering, Dhankawadi, Pune, 411 043 India
shobha.raskar@gmail.com

Prof. D. M. Thakore
Bharati Vidyapeeth University, BVUCOE, Department of computer Science and Engineering, Dhankawadi, Pune, 411043 India.

***Abstract:*** Text mining is process of finding meaningful information from large amount of unstructured text documents. Key phrases are an important means of document summarization, clustering, and topic search. Only a small minority of documents have author-assigned keyphrases, and manually assigning keyphrases to existing documents is very tedious. Therefore it is highly desirable to automate the keyphrase extraction process. Kea-mean clustering Algorithm is combination of k-mean and keyphrase extraction algorithm. In Kea-means algorithm, documents are clustered into several groups like K-means, but the number of clusters is determined automatically by using the extracted keyphrases.
Set of training documents and machine learning is used to determine phrases are keyphrase or not. Cluster analysis is required in text mining for grouping objects. Keyphrase extraction algorithm returns several keyphrases from the source documents. The Kea-means clustering algorithm provides easy and efficient way to extract documents from document resources.

***Keywords:*** Text mining, Keyphrase extraction, clustering, Categorization, Neural Networks

## I. INTRODUCTION

Text mining is a mechanism to understand and extract meaningful information from large amount of the semi-structured or unstructured text data. Information extraction identifies Keyphrase and relationships within text. A keyphrase is defined as meaningful and significant expression consisting of one or more words in documents [3]. Set of training documents and machine learning is used to determine phrases are keyphrase or not.

The learning process is use to find a mapping from documents to categories using a set of training documents, which can be accomplished by training a classifier for each category. For Machine Learning, Artificial Neural Network is used. In supervised learning the network user assembles a set of *training data*. The training data contains examples of inputs together with the corresponding outputs, and the network learns to understand the relationship between the two**.**

A new document is then processed by each of the classifiers and assigned to those categories whose classifiers identify it as a positive example. Cluster analysis is required in text mining for grouping objects It does this by looking for predefined sequences in text, a process called pattern matching. These procedures contain text summarization, text categorization, and text clustering. Text summarization is the procedure to extract its partial content reflection its whole contents automatically. Text categorization is the procedure of assigning a category to the text among categories predefined by users. Text clustering is the procedure of segmenting texts into several clusters, depending on the substantial relevance.

### *A. Pre-processing steps for Text:*

The first step in text clustering is to transform documents, which typically are strings of characters into a suitable representation for the clustering task. Text data are pre-processed with the aid of stop words removal technique and stemming algorithm [1].

***a. Remove stop-words:*** The stop-words are high frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc.). Remove stop-words can improve clustering results.


Figure 1: Pre-processing Steps

***b. Stemming:*** **S**temming means the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, such as work, worker, worked and working.

## II. ABOUT CLUSTERING

The "means" in k-means refers to the centroid of the cluster, which is a data point that is chosen arbitrarily and

then refined iteratively until it represents the true mean of all data points in the cluster. The "k" refers to an arbitrary number of points that are used to seed the clustering process. The k-means algorithm assigns each data point to exactly one cluster. Text clustering process contains four main parts: text reprocessing, word relativity computation, word clustering and text classification. The first step in text clustering is to transform documents, which typically are strings of characters into a suitable representation for the clustering task.

***Step 1:*** Place randomly initial group centroids into the two dimension space.

***Step 2:*** Assign each object to the group that has the closest centroid.

***Step 3***: Recalculate the positions of the centroids.

***Step 4:*** If the positions of the centroids didn't change go to the next step, else go to Step 2.

***Step 5:*** End.

### *A. About kea-means clustering:*

The Kea-means clustering algorithm used for clustering that improves the K-means algorithm by combining it with the keyphrase extraction algorithm. Main drawback of K-means clustering that the number of total clusters is pre-specified in advance. The Kea-means clustering tries to solve the main drawback of K-means clustering.

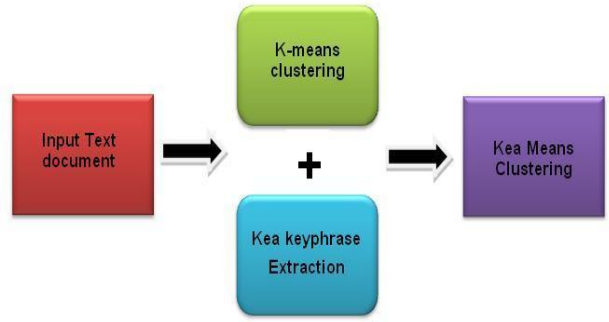


Figure 2: kea-mean clustering

In Kea-means algorithm, documents are clustered into several groups like K-means, but the number of clusters is determined automatically by the algorithm by using the extracted keyphrases. The system architecture of the Kea-means clustering is shown in Figure 2.

## III. ABOUT KEYPHRASE

A keyphrase is "a sequence of one or more words that is considered highly relevant". Keyphrase provide information about documents. Keyphrases give a high-level description of a document's contents that decide whether or not it is relevant for them. Entering a keyphrase into a search engine, all documents with this particular keyphrase attached are returned to the user. Manually attaching keyphrases to existing documents is a very laborious task. Therefore automatic keyphrase extraction is used. There are two different ways of approaching the problem:

a) keyphrase assignment
b) keyphrase extraction

In keyphrase assignment, it is assumed that all potential keyphrases appear in a predefined controlled vocabulary. The learning problem is to find a mapping from documents to categories using a set of training documents. A new document is then processed by each of the classifiers and assigned to those categories whose classifiers identify it as a positive example.

## IV. ABOUT CATEGORIZATION AND CENTROID BASED CLASSIFIER

The text categorization task is to train the classifier using documents, and assign categories to new documents.

### *A. Centroid-Based Classifier:*

a. Input: new document *d;*
b. Training collection:D={$d_1$,$d_2$,…$d_n$ };
c. predefined categories:C={$c_1$,$c_2$,….,$c_l$};
d. //Compute similarities
for($d_i$∈D){ Simil(*d*,$d_i$) =cos(*d*,$d_i$); }
e. //Select k-nearest neighbour

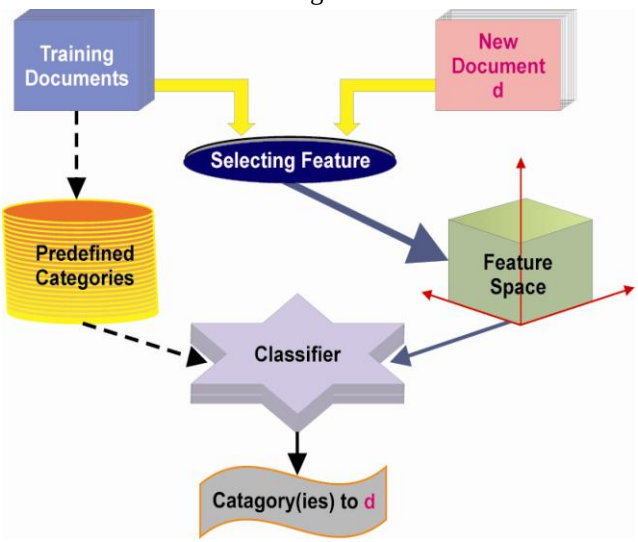


Figure 3: Categorization Architecture

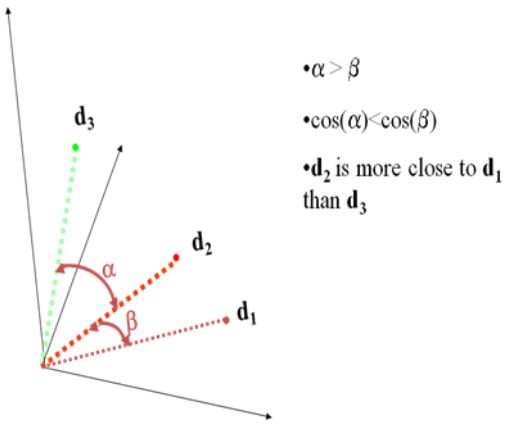


Figure 4: Model: K-Nearest Neighbour Classifier

Construct k-document subset $D_k$ so that
Simil(*d*,$d_i$) < min(Simil(*d*,doc)

f. //Compute score for each category
g. //Output: Assign to *d* the category *c with the highest score:*

## V. BUILDING MODEL: TRAINING AND TESTING

Kea's extraction algorithm has two stages, training and extraction. The training stage uses a set of training documents for which the author's keyphrases are known.

Each training document, candidate phrases are identified and their feature values are calculated. For each phrase is then marked as a keyphrase or a non-keyphrase, using the actual keyphrases for that document.

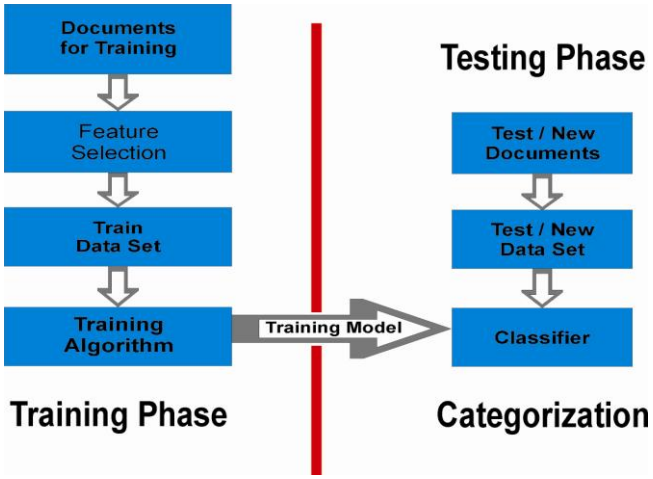


Figure 5 : Training and Testing

The text categorization task is to train the classifier using these documents, and assign categories to new documents. In the training phase, the *n* documents are arranged in *p* separate folders, where each folder corresponds to one class. In the next step, the training data set is prepared via a feature selection process.

Training is a process for making the machine, learn something from the environment by experience. Learning is an inherent characteristic of the human beings. When this learning is done by a machine, it is usually referred to as 'machine learning'.

The training stage uses a set of training documents for which the author's keyphrases are known. For each training document, candidate phrases are identified and their feature values are calculated. Each phrase is then marked as a keyphrase or a nonkeyphrase, using the actual keyphrases for that document.

Supervised learning requires a trainer, who supplies the input-output training instances. The learning system adapts its parameters by some algorithms to generate the desired output patterns from a given input pattern. The input data for the extractor and also the model builder has to be in text files with the ending .txt and all in the same directory. After removing stopwords next step to find a set of candidates for phrases. Finally stem of the candidates are searched with the help from Lovins stemmer. Feature value calculated. Keyphrase extraction model build by training the system with texts files and their key phrases. These training documents must have the ending '.txt'. Their keyphrase documents are also text documents with the same name but with the ending '.key'

Use this model for extracting key phrases from more documents. The model determines the overall probability that each candidate is a keyphrase, and then a post-processing operation selects the best set of keyphrases. To select keyphrases from a new document, Kea determines candidate phrases and feature values, and then applies the model built during training. The model determines the overall probability that each candidate is a keyphrase, and then a post-processing operation selects the best set of keyphrases.

| | A | B | K | O | Q | R | S | T | W | X |
|---|---|---|---|---|---|---|---|---|---|---|
| D1 | 3 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| D2 | 3 | 2 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |

Figure 6 : Weighting model: example

$$W_{ij} = Freq_{ij} * \log(N/ DocFreqj) \quad (1)$$

Where:

a. N = the number of documents in the training document collection.
b. tf = Term Frequency weighting
c. $DocFreq_j$= the number of documents in which the $j^{th}$ term occurs
d. Freqij= the number of times $j^{th}$ term occurs in document

### *A. Extracting Keyphrases:*

Kea uses set of training documents to generate model for which keyphrases are known. The resulting model can then be applied to a new document from which keyphrases are to be extracted. First, Kea computes TFIDF scores and distance values for all phrases in the new document.

## VI. ABOUT MACHINE LEARNING: ANN

Artificial neural network (ANN) is set of nodes (units, neurons, processing elements)

a. Each node has input and output
b. Each node performs a simple computation by its node function

ANN refer to multilayer perceptron (MLP) network, which is most widely used type of neural network [4]. It consists of input layer, one or more hidden layers and output layers.

***Input:*** D is dataset consisting of training tuples and their associated target values.

***Output:*** A trained neural network

Weights in network are initialize to small random numbers.

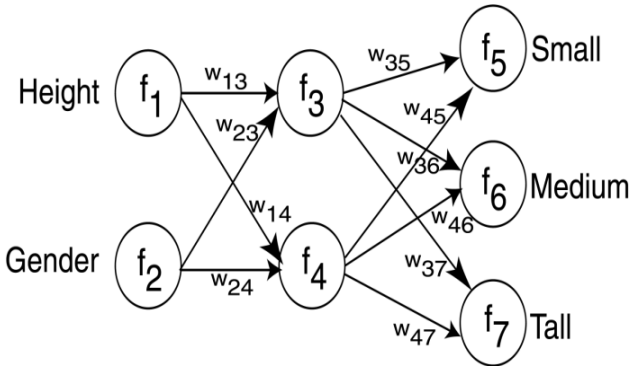


Figure 7: ANN model example

A Neural Network Model is a computational model consisting of three parts:

a) Neural Network graph
b) Learning algorithm that indicates how learning takes place.
c) Recall techniques that determine how information is obtained from the network.

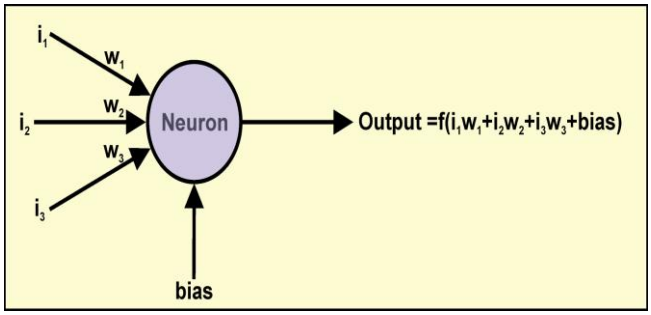


Figure 8: Neural Network node

The artificial neural network consists of many artificial neurons. An artificial neuron is roughly imitating brain neuron, and it receives information from outside or other artificial neurons. It is use of easy operation, and output results to outside or other artificial neurons. In artificial neural model, an artificial neuron is called processing unit, outputs of every processing unit send out in fan shape, become inputs of other processing unit. The relation between the inputs and outputs of a processing unit is representatives of sum of weighted product function, as shown in figure 9.

## VII. RESULTS

System implemented in JDK1.6 and input dataset consists of text files containing only text data. To extract keyphrases from documents first of all model built which can be used for the purpose of extraction, and for this system is train from some known facts using supervised learning. Hence the implementation of algorithm consists of two steps, Training and extraction. Training is a process for making the machine, learn something from the environment by experience. When learning is done by a machine, it is usually referred to as 'machine learning'. Supervised learning requires a trainer, who supplies the input-output training instances. The learning system adapts its parameters by some algorithms to generate the desired output patterns from a given input pattern.

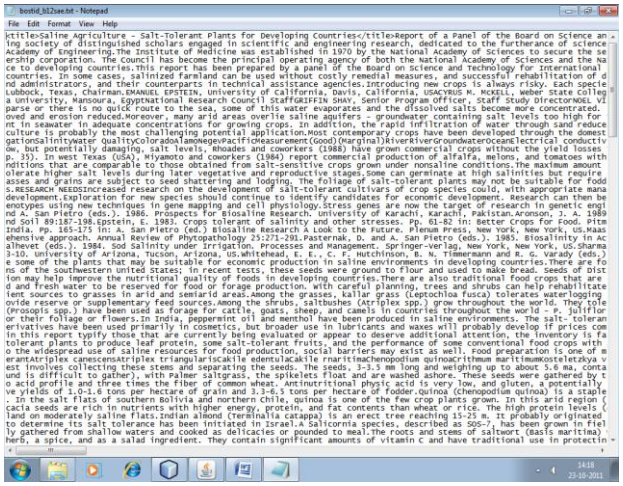

Figure 9: Example '.txt' file

The input data for the extractor and also the model builder has to be in text files with the ending .txt and all in the same directory. The text should be as clean as possible i.e. without code tags. The language of the input should be English. After removing stopwords next step to find a set of candidates for phrases. finally stem of the candidates are searched with the help from Lovins stemmer. Feature value calculated.

### *A. Experiment I:*

The keyphrases discovered are stored in a '.key' file for each test file. In table 1, there are only three equal terms in the two columns so algorithm discovers three words of the original author keyphrase phrases. But, the results obtained are not bad, because the other keyphrase phrases discovered are related to the original ones and to the topic of the course.

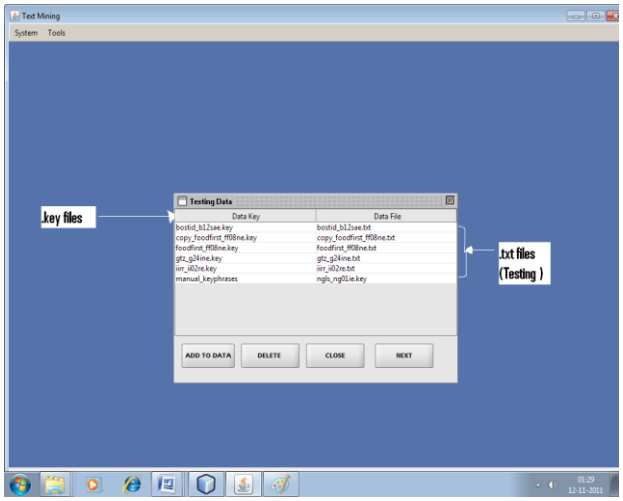


Figure : 10 Adding and deleting testing phase data

Table 1. Keyphrases assigned by author and extracted by KEA.

| Author keyphrases | Kea keyphrases |
|---|---|
| addition | addition |
| Loop | Exercise |
| condition | condition |
| Variable | Error |
| Value | Value |
| Procedure | Runtime |

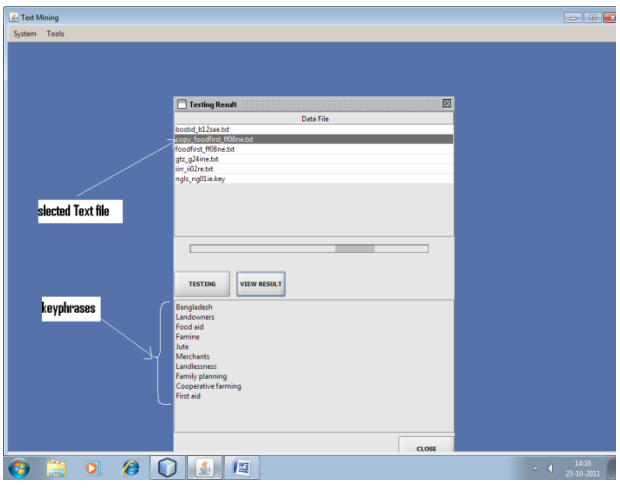


Figure 11: Result showing keyphrases of selected text documents

### *B. Experiment II:*

In clustering statistical classification includes precision and recall. Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance. Table 2 gives the difference between Precision and Recall.

Table 2 : Comparison of Precision and Recall

| Precision | Recall |
|---|---|
| 1) It is measure of extractness. | 1) It is measure of |
| 2) Precision is equal to number of relevant document retrieved by search divided by total number of documents retrieved by that search | 2) Recall is equal to number of relevant document retrieved by search divided by total number of documents. |
| 3) It measure how precise search is. | 3) It measure how complete search is. |
| 4) Higher precision means less unwanted documents. | 4) Higher recall means less missing documents. |

There are five test documents are used. The result of each of the individual test document for top 3, top 5 keyphrases is shown by using the proposed system.

Table 3: Comparison of Proposed System Precision and Recall

| Document No | Proposed System | | | | |
|---|---|---|---|---|---|
| | TP | FP | FN | Prec | Rec |
| Doc1(Top-3) | 1 | 2 | 8 | 0.33 | 0.11 |
| Doc1(Top-5) | 3 | 2 | 6 | 0.6 | 0.33 |
| | | | | | |
| **Document No** | **Proposed System** | | | | |
| | TP | FP | FN | Prec | Rec |
| Doc2(Top-3) | 2 | 1 | 8 | 0.66 | 0.2 |
| Doc2(Top-5) | 3 | 2 | 7 | 0.6 | 0.3 |
| | | | | | |
| **Document No** | **Proposed System** | | | | |
| | TP | FP | FN | Prec | Rec |
| Doc3(Top-3) | 2 | 1 | 2 | 0.67 | 0.5 |
| Doc3(Top-5) | 2 | 3 | 2 | 0.4 | 0.5 |
| | | | | | |
| **Document No** | **Proposed System** | | | | |
| | TP | FP | FN | Prec | Rec |
| Doc4(Top-3) | 1 | 2 | 6 | 0.33 | 0.14 |
| Doc4(Top-5) | 2 | 3 | 5 | 0.4 | 0.29 |
| | | | | | |
| **Document No** | **Proposed System** | | | | |
| | TP | FP | FN | Prec | Rec |
| Doc5(Top-3) | 1 | 2 | 16 | 0.33 | 0.06 |
| Doc5(Top-5) | 1 | 4 | 16 | 0.2 | 0.06 |

***a) Precision (P):*** The percentage of correctly extracted tags out of all extracted

***b) Recall(R):*** The percentage of correct extracted tags out of all correct

$$P = \frac{TP}{TP + FP} \quad (2)$$

$$R = \frac{TP}{TP + FN} \quad (3)$$

Where,

a. TP = True Positive, (correct result)
b. FP = False Positive,(unexpected result)
c. FN = False Negative,(missing result)

The result of extraction of keyphrases depends very much on the domain of the training set of documents, because these are the training instances that make the system learn. To get better result of extraction of keypharses from any business text document, it is necessary to train the system on some documents of the same domain.

## VIII. CONCLUSION AND FUTURE WORK

Keyphrase and K-mean clustering algorithm is important for obtaining the appropriate cluster context and the low quality clustering results will decrease extraction performance. Kea mean algorithm provides efficient way to extract test documents from large quantity of resources. This algorithm develop faster algorithm for clustering. To get better result of extraction, it is necessary to train the system on same domain. In the future, need to develop a faster algorithm for clustering.

## IX. REFERENCES

[1] S. Murali Krishna, S. Durga Bhavani, "An efficient approach for Text Clustering Based on Frequent Itemsets",Europen Jounal of Scientific Research, ISSN 1450-216X Vol. 42 No.3 (2010) PP 385-396

[2] N. Kanya, S. Geetha, "Information Extraction-A text Mining Approach", IET-UK International Conference on Information and communication Technology inEletrical Science (ICTES 2007), pp 1111-118.

[3] Xiaojun Wan and Jianguo Xiao, "CollabRank: Towards a Collaborative Approach to Single document Keyphrase Extraction", Proceeding of 22nd Intenational Conference on Computational Linguistics , Aug 2008 , PP 969-976.

[4] Kamal Sarkar, Mita N, Suranjan G, " A New Approach to Keyphrase Extraction Using Neural Networks", IJCSI, Vol 7, Issue 2. No.1 , March 2010, ISSN 1694-0784.

[5] D. Sanchez, M. J. M, I. Balanco, " Text Knowledge Mining : An Alternative to Text Data Mining ", 2008 IEEE International Conference on Data Mining Workshops.

[6] Decong Li, Sujian Li, Wenjie Li, Wei Wang, Weiguang Qu, "A Semi-Supervised Key Phrase Extraction Approach: Learning from Title Phrases through a Document Semantic Network" , Proceedings of the ACL 2010 Conference Short Papers, pages 296–300, Uppsala, Sweden, 11-16 July 2010, Association for Computational Linguistics.

**Short Biodata of Authors**

**Mrs. Shobha S. Raskar**, is a student of M. Tech in the Department of Computer Engineering, Bharati Vidyapeeth University College of Engineering, Pune. She obtained her B.E. Computer Engineering, from Cummins College of Engineering for Women, Pune University. Her research interest is, Text mining/Data mining.

**Prof. D. M. Thakore** graduated (B.E–Computer Engineering) from Walchand College of Engineering, Sangli and State–Maharashtra in 1990. He pursued his M.E. (Computer) from Bharati Vidyapeeth University College of Engineering, Pune in 2004. He also completed MBA (Marketing) from Pune University in 1996. He is currently working and pursuing his Ph.D. with subject Data Mining/Text Mining from Bharati Vidyapeeth Deemed University College of Engineering, Pune