A Novel Multi label Text Classification Model using Semi supervised learning

Shweta C. Dharmadhikari, Maya Ingle, Parag Kulkarni

Abstract:

Automatic text categorization (ATC) is a prominent research area within Information retrieval. Through this paper a classification model for ATC in multi-label domain is discussed. We are proposing a new multi label text classification model for assigning more relevant set of categories to every input text document. Our model is greatly influenced by graph based framework and Semi supervised learning. We demonstrate the effectiveness of our model using Enron , Slashdot , Bibtex and RCV1 datasets. Our experimental results indicate that the use of Semi Supervised Learning in MLTC greatly improves the decision making capability of classifier.

Keywords:

Automatic text categorization, Multi-label text classification, graph based framework , semi supervised learning.

1. INTRODUCTION

The amount of textual data being produced through internet is growing faster than the ability of information consumers to search, digest and use it. Textual data is difficult to effectively understand and categorize because the relationship between its sequence of words and its content is less clear as compared to numerical. Such data includes technical article, memos, manuals, electronic mail, books, online news paper, journal articles and many other forms of texts. Thus text classification has become an active research topic now a day. It classifies document under a predefined category. Categories may be represented numerically or using single word or phrase or words with senses, etc. In traditional approach, classification of text was carried out manually using domain experts. The human expert was required to read and sort the input text document to predefined category or set of categories. Thus this approach requires extensive human efforts and error prone also. This leads to the scheme of automated text classification scenario. This automated text document classification facilitates ease of storage, searching, retrieval of relevant text documents or its contents for the needy applications. Three different paradigm exists under text classification and they are single label(Binary) , multiclass and multi label. Under single label a new text document belongs to exactly one of two given classes, in multi-class case a new text document belongs to just one class of a set of m classes and under multi label text classification scheme each document may belong to several classes simultaneously [3]. In real practice many approaches are exists and proposed for binary case and multi class case even though in many applications text documents are inherently multi label in nature. Eg. In medical diagnosis a document report containing set of symptoms can belong to many probable disease categories. Multilabel text classification problem refers to the scenario in which a text document can be assigned to more than one classes simultaneously during the process of classification. Eg. In the

DOI : 10.5121/ijdkp.2012.2402

process of classification of online news article the news stories about the scams in the commonwealth games in india can belong to classes like sports, politics , country-india etc. It has attracted significant attention from lot of researchers for playing crucial role in many applications such as web page classification, classification of news articles , information retrieval etc.

Many approaches are existing to implement multi label text classifier. Supervised methods from machine learning are more popular amongst these. But majority of existing approaches are lacking in considering relationship between class labels, input documents and also relying on labeled data all the time for classification. In real life unlabeled data is readily available whereas generation of labeled data is expensive and error prone as it needs human intervention. In many situations the available class labels are related to each other and consideration of this relationship can lead to better accuracy. Also, the abundantly available unlabeled data contains the joint distribution over features of a input dataset which may improve accuracy of overall classification process when used in conjunction with labeled data. So in our proposed classification model we are considering the class correlations and semi supervised learning scheme to learn the classifier to overcome the limitations of existing approaches. We are also trying to remove redundant data from input dataset as it affects performance of classifier.

All the existing approaches needs initial step of text document representation[16]. The common approaches are vector space model using various term weighting schemes such as Boolean , word frequency count , term and document frequency , entropy encoding etc. All of these are popularly known as BOW (Bag Of Words) approaches[17]. Even though these are widely used but these ignores use of structural and semantic information in classification which may significantly improves accuracy. Other alternative to bag of words representation is graph based representation. The graph based representation offers much better document representation as it also considers relationship among documents in the form of edge of the graph[16].

Through our paper we are proposing a classification model which is exploiting relationship between input and class labels as a graph with the setting of semi supervised learning to use unlabeled data effectively for classification along with labeled data. Through this set up we are aiming at improving decision making capacity of multi label text classifier. We apply the proposed framework on standard dataset such as Enron, Bibtex and RCV1 and Slashdot to test the performance.

The rest of the paper is organized as below. Section 2 describes literature related to construction of multi label text classification system ; Section 3 highlights overview of graph representation . Section 4 describes our proposed classification model followed by experiments and results in Section 5 , followed by a conclusion in the last section.

2. RELATED WORK / LITERATURE

Multi label text classifier can be realized by using supervised, unsupervised and semi supervised methods of machine learning. In supervised methods only labeled text data is needed for training. Unsupervised methods relies heavily on only unlabeled text documents; whereas semi supervised methods can effectively use unlabeled data in addition to the labeled data[1][2].

The most traditional approach towards multi-label learning decomposes the classification task into multiple independent binary classification tasks, one for each category. But its major

drawback is that it can not scale to a large number of class labels and does not exploit relationship among class labels while predicting class labels of test documents[6]. Another general method is to learn the ranking function of category labels from the labeled instances and apply it to classify each unknown test instance by choosing all the categories with the scores above the given threshold. But these methods also do not exploit relationship among class labels[10]. Few other popular existing methods are binary relevance method, label power set method, pruned sets method, C4.5, Adaboost.MH & Adaboost.MR, ML-kNN , Classifier chains method etc[20]. But all these are lacking the capability of handling unlabeled data ie these are based on principle of supervised learning and these can not exploit class relationships.

Recently some new approaches for multi-label learning that consider the correlations among categories have been developed. Few eg. are generative model proposed by Ueda[26] , Bayesian model proposed by Griffiths [27] , Hierarchical structure considered by Rousu [28] , Maximum entropy method proposed by Zhu[29] , Latent variable based approach proposed by McCallum. But all these methods are also supervised in nature.

Traditional graph based semi-supervised methods only construct a graph at input instance level. It gives good results when there are no correlations among categories. But in many practical situations , there often exists relation among category labels. Therefore, in order to make use of the correlation information, we have not only constructed graph at input instance level but at category level also.

While designing a multi label text classifier the major objective is not only to identify the set of classes belonging to given new text documents but also to identify most relevant out of them to improve accuracy of overall classification process. Graph based approaches are known for their effective exploration of document representation and semi supervised methods explores both labeled and unlabeled data for classification that's why accuracy of multi label text classifier can be improved by using graph based representation of input documents and class labels in conjunction with label propagation approach of semi supervised learning[16][17].

Few approaches are proposed based on the combination of graph representation and semi-supervised learning. In 2006 Liu, Jin and Yan proposed Multi-label classification approach based on constrained non negative matrix factorization [8]. In this approach parameter selection affects the overall performance of the system. Zha and Mie proposed Graph-based SSL for multi-label classification in the year 2008[9]. But this approach was purely intended for classification of video files and not for documents. Chen,Song and Zhang proposed Semi supervised multi-label learning by solving a Sylvester Eq in the year 2010 [10]. In this approach they constructed graph for input representation and class representation as well but this approach is getting slower on convergence when applied in the situation where large number of classes and input data exists. In 2009 Lee, Yoo and Choi proposed Semi-Supervised Non negative Matrix Factorization based approach [11]. But this approach was not specifically meant for multi-label text classification. Through our model we are proposing semi-supervised learning based multi-label text classification model in which graph based framework is employed in preprocessing step to improve the accuracy.

In our proposed model ,preprocessing stage exploits relationship between labeled and unlabeled documents by identifying structural and semantically relationship between them for more relevant classification through graph.; and during training stage semi supervised methods are used to propagate labels of labeled documents to unlabeled documents based on some energy function.

3. MATHEMATICAL MODEL OF PROPOSED SYSTEM

Our overall multi label classification system **S** is defined as follows:

S = <D, C, T, Ø> where 'D' represents set of document corpus, 'C' represents set of classes, 'T' represents set of training set consisting of <document, class label set > pair.

'Ø' is multi label assignment function used to predict the set of labels for unlabeled documents $\forall$ **Ø: D** $\rightarrow$ $\mathbf{2^c}$.

Document corpus D is represented as $D = \{d_1,\ldots, d_n\}$, where n is the total number of documents in the document corpus. Out of these "l" no. of documents are labeled and remaining are unlabeled and represented by "u". Thus $n = l + u$.

Every document d_i in turn represented as m – dimensional feature vector and represented as: $d_i = (d_{i1}, \ldots d_{im})$.

Similarly $C = \{ C_1, \ldots . , C_n \}$ represents set of classes. Each of the class label represents set of classes , $Ci=<c_1,..,c_m>$.

T represents multi label training set as, $T = \{(d_1, C_1), (d_2, C_2), \ldots, (d_l, C_l)\}$.

G(V,E) represents a connected graph, where V represents set of vertices corresponding to input document corpus. Thus total no. of vertices of graph equals to 'n', out of that 'l' no. of vertices are labeled and 'u' no. of vertices are unlabelled. The objective is to predict the labels of nodes 'u'.

4. PROPOSED ARCHITECTURE OF CLASSIFIER MODEL

Our proposed classifier architecture works in two phases namely training phase and testing phase. We have used labeled as well as unlabeled data for training. These two phases are depicted in fig. 1 and works as follows:

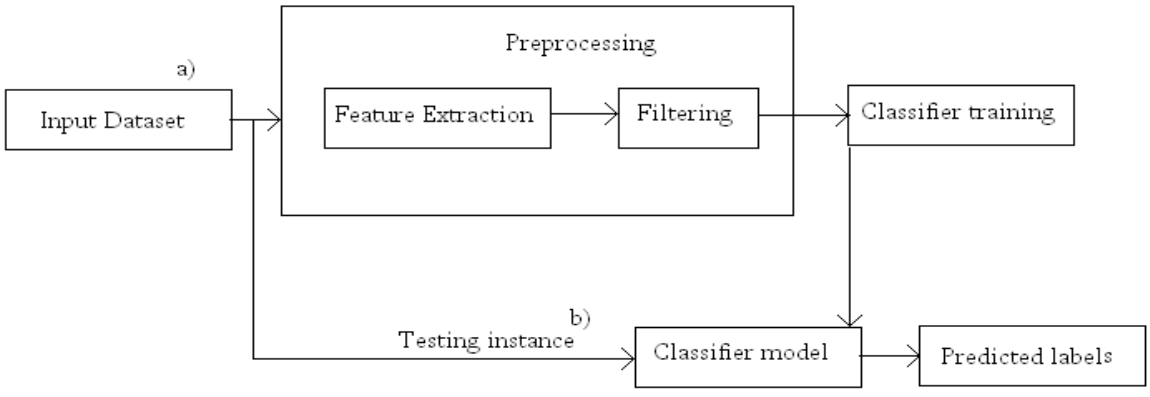

Fig. 1 Architecture of Classifier Model

3.1) Training phase: From the fig.1 a) part represents training phase. Preprocessed data is given as input to this phase.

Preprocessing: This phase converts the input dataset into the form of weighted graph. The input data set is represented in the form of vector space model. It uses TF * IDF measure for every term. It constitutes of feature extraction and filtering stage.

Feature extraction: To this sub stage, input is input document corpus, D. |D|=n i.e. n=l+u . as per mathematical model description. So given input as D, this phase constructs a graph G(V,E) with |V|=n. Each vertex Vi represents document instance d_i . Relationship between pair of vertices is represented by edge E. The adjacency matrix $A \in Rnxn$ is computed to represent the edge weight using cosine similarity measures. We have captured the correlation among different classes by computing matrix [B]kxk for representing relationship between classes.

Filtering: In this, Matrix A is sparcified and reweighed using K-nn approach and produce matrix W. $A \Rightarrow W \in Rnxn$. This graph sparcification can lead to improved efficiency in the label inference stage.

3.1.2) Classifier training: To this phase specified graph W acts as an input. Given this graph W and label information. This phase infers labels of unlabeled documents. It estimates a continuous classification function F on W i.e.

$F \in R_{|v|x|c|}$ Where |v| is number of vertices and |c| is number of class labels.

$F: W \rightarrow \hat{C}_U$ …Where $\hat{C}_U$ is estimated label set for unclassified document.

It estimates soft labels of unlabeled doc. By optimizing the energy function by generating confidence matrix [P]nxn.

3.2) Prediction phase: Prediction is made by classifier model generated by training phase. Our prediction policy works on the smoothness assumption of SSL which states that "If two input points x1, x2 are in high density region are closer to each other then so should be the corresponding outputs y1,y2". Closeness between the two document instance can be identified by W. Relation between corresponding class labels can be computed by weighted dot product piBpj . If assignment of class labels pi and pj are relevant to doc. di and dj then we would expect $W_{i,j} \approx$ piBpj and uses following smoothness function to predict the labels of unlabeled doc.

$$\emptyset = \sum_{i,j=1}^{n} \left(W_{i,j} - \sum_{i,k=1}^{m} p_i B p_j \right)$$

5. WORKING OF PROPOSED MODEL

Input: $D=\{d_1,\ldots d_n\}$, T

Output: Labels C^U (Labels for unlabeled doc. instances)

1. Represent input doc. Corpus D in vector space model $D \Rightarrow V$

   $\forall\ d_j = (W_{ij}, W_{2j}, \ldots.., W_{tj})$

   Where W_{ij} is the weight of the word i in the doc. j and computed by tf|Df

2. Constructs a weighted undirected graph $G(V \Rightarrow G)$ represented as a adjacency matrix $[A]_{nxn}$

3. Sparcify a reweighed G for noise removal
   $A \Rightarrow W \in R^{nxn}$

4. Train the classifier, define energy function F to propagate soft labels to unlabeled doc. and
   F: $W \rightarrow \hat{C}_U$
   Where $\hat{C}$ is estimated labels generate confidence score matrix

5. Predict the class labels using smoothness function
   $\emptyset = \sum_{i,j=1}^{n} \left(W_{i,j} - \sum_{i,k=1}^{m} p_i B p_j \right)$

6. EXPERIMENTATIONS AND RESULT DISCUSSION

In order to evaluate the performance of our proposed classification approach, we conducted experiments in order to-

- Investigate the performance of our classification model based on semi supervised learning in terms of accuracy, F-measure, precision and recall .
- Investigate the performance of our proposed model against few popular supervised methods for multi label text classification.

We evaluated our approach under a WEKA-based [23] framework running under Java JDK 1.6 with the libraries of MEKA and Mulan [21][22]. Jblas library for performing matrix operations while computing weights on graph edges. Experiments ran on 64 bit machines with 2.6 GHz of clock speed, allowing up to 4 GB RAM per iteration. Ensemble iterations are set to 10 for EPS. Evaluation is done in the form of 5 × 2 fold cross validation on each dataset .We first measured the accuracy, precision, Recall after label propagation phase is over.

6.1 Datasets used for Experimentations

We tested the performance of our proposed model on four bench mark datasets namely Enron , Slashdot , Bibtex and Reuters. Table I summarizes the statistics of datasets that we used in our experiments.

TABLE I : STATISTICS OF DATASETS

Dataset	No. of document instances	No. of Labels	Attributes
Slashdot	3782	22	500
Enron	1702	53	1001
Bibtex	7395	159	1836

Enron dataset contains email messages. It is a subset of about 1700 labeled email messages[21]. BibTeX data set contains metadata for the bibtex items like the title of the paper, the authors, etc. Slashdot dataset contains article titles and partial blurbs mined from Slashdot.org[22].

We first measured the accuracy, precision ,Recall and after label propagation phase is over. Figure 2 shows comparison of accuracy measured for each dataset. In order to evaluate the performance of our classifier model using SSL approach, we compared the results of few popular supervised algorithm such as Binary Relevance (BR),C4.5, SVM-HF, Classifier chains method(CC), Pruned Sets Method (PS) and our proposed approach (referred as lblMLTC). Fig. 3 represents comparison of accuracy measured between our lblMLTC and other supervised approach on the same benchmark datasets.

We used accuracy measure proposed by Godbole and Sarawagi in [13] . It symmetrically measures how close y_i is to Zi ie estimated labels and true labels. It is the ratio of the size of the union and intersection of the predicted and actual label sets, taken for each example and averaged over the number of examples. The formula used by them to compute accuracy is as follows:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{Y_i \cap Z_i}{Y_i \cup Z_i}\right]$$

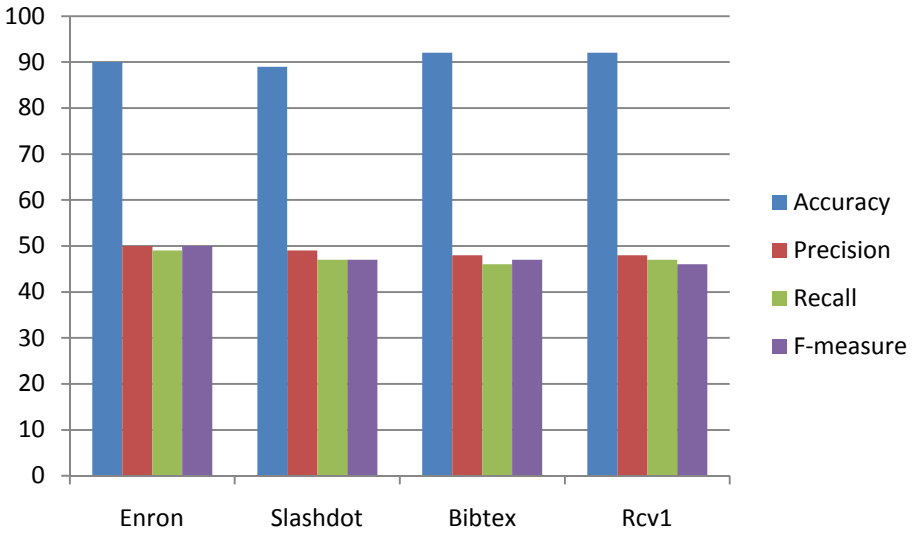

Fig. 2 Comparison of results measured using lblMLTC on four benchmark datasets

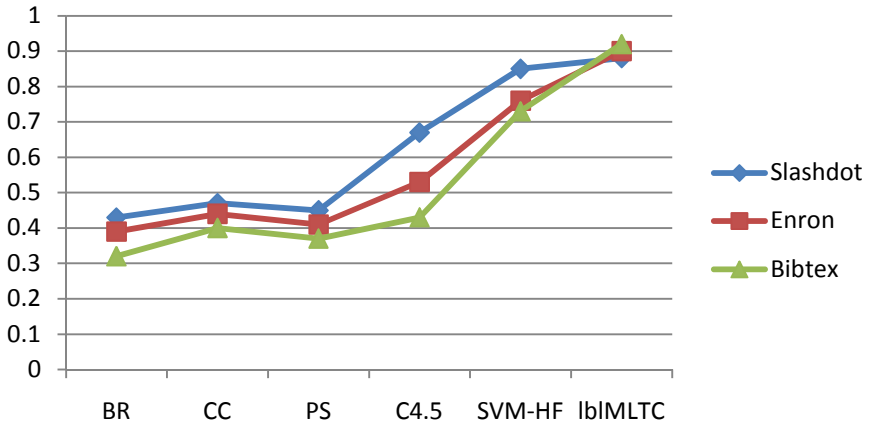

Fig. 3 comparison of accuracy measured using lblMLTC and other popular supervised approaches.

7. CONCLUSION

A new multi-label text classification model using a graph based representation and semi supervised learning has been described. In our classification model we incorporated document similarity along with class label correlation in order to improve accuracy of multi label text classifier.We have used semi-supervised learning to utilize the unlabeled data for text classification.We have evaluated our classification model against small scale as well as large scale datasets. Experimental results show that our model offers reasonably good accuracy. Empirical studies show that our approach is quite competitive against supervised multi-label text classification techniques. Use of cosine similarity measure may ignore some aspects of semantic relationship between text documents which can affect accuracy. However In future, along with vector space model of text representation use of more robust feature extraction technique like LSI or NMF may be incorporated in order to reduce rate of misclassification.

REFERENCES

[1] J. Zhu. Semi-supervised learning Literature Survey. Computer Science Technical Report TR 1530, University of Wisconsin – Madison , 2005.

[2] Olivier Chapelle, Bernhard Schfolkopf, Alexander Zien. Semi-Supervised Learning 2006, 03-08, MIT Press.

[3] G. Tsoumakas, I. Katakis. Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 3(3):1-13, 2007.

[4] A. Santos, A.Canuto, and A.Neto, “A comparative analysis of classification methods to multi-label tasks in different application domains”, International journal of computer Information systems and Industrial Management Applications”. ISSN: 2150-7988 volume 3(2011), pp. 218-227.

[5] R.Cerri, R.R. Silva, and A.C. Carvalho, “Comparing Methods for multilabel classification of proteins using machine learning techniques”, BSB 2009, LNCS 5676,109-120,2009.

[6] G. Tsoumakas, G.Kalliris and I. Vlahavas, " Effective and efficient multilabel classification in domains with large number of labels", Proc.Of the ECML/PKDD 2008 workshop on Mining Multidimensional Data (MMD' 08)(2008) 30-44.

[7] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. M. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning,39, 103–134.

[8] Y. Liu, R. Jin, L. Yang. Semi-supervised Multi-label Learning by Constrained Non-Negative Matrix Factorization .In: AAAI, 2006.

[9] Z. Zha, T. Mie, Z. Wang, X. Hua. Graph-Based Semi-Supervised Learning with Multi-label. In ICME. page 1321-1324, 2008.

[10] G. Chen, Y. Song, C. Zhang. Semi-supervised Multi-label Learning by Solving a Sylvester Equation. In SDM, 2008.

[11] Semi-supervised Nonnegative Matrix factorization. IEEE. January 2011.

[12] Qu Wei , Yang, Junping, Wang. Semi-supervised Multi- label Learning Algorithm using dependency among labels. In IPCSIT vol. 3 2011.

[13] S.Godbole and S. Sarawagi , "Discriminative methods for multi-labeled classification", 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2004.

[14] R. Angelova, G.Weikum . "Graph based text classification : Learn from your neighbours". In SIGIR'06, ACM , 1-59593-369-7/06/0008".

[15] T.Jebara , Wang and chang , "Graph construction and b-matching for semi supervised learning". In proceedings of ICML-2009.

[16] Thomas, Ilias & Nello. " scalable corpus annotation by graph construction and label propogation". In proceedings of ICPRAM, 25-34, 2012.

[17] P. Talukdar , F. Pereira. " Experimentation in graph based semi supervised learning methods for class instance acquisition". In the proceedings of 48th Annual meet of ACL. 1473-1481.2010.

[18] X. Dai, B. Tian , J. Zhou , J. Chen. "Incorporating LSI into spectral graph transducer for text classification". In the proceedings of AAAI. 2008.

[19] S.C.Dharmadhikari, Maya Ingle, parag Kulkarni .Analysis of semi supervised methods towards multi-label text classification. IJCA , Vol. 42, pp. 15-20 ISBN :973-93-80866-84-5.

[20] S.C.Dharmadhikari, Maya Ingle, parag Kulkarni .A comparative analysis of supervised multi-label text classification methods. IJERA , Vol. 1, Issue 4, pp. 1952-1961 ISSN : 2248-9622.

[21] http://mulan.sourceforge.net/datasets.html

[22] http://MEKA.sourceforge.net

[23] www.cs.waikato.ac.nz/ml/weka/

[24] J. Read, B. Pfahringer, G. Homes and E.Frank , "Classifier chains for multi-label classification"., Proc. Of European Conference on Machine Learning and knowledge discovery in Databases, LNAI 5782(254-269),2009.

[25] R. E. Schapire and Y. Singer. Boostexter: A boosting based system for text categorization. Machine learning, 39(2-3),2000.

[26] Ueda and K. Saito. Parametric mixture models for multi-labelled text. In proc. of NIPS, 2002.

[27] Griffiths and Ghahramani. Infinite latent feature models and the Indian buffet process. In proc. of NIPS, 2005.

[28] Rousu, saunders. On maximum margin hierarchical multi-label classification. In Proc. Of NIPS workshop on Learning with structured outputs ,2004.

[29] S. Zhu, Ji, Xu and Y. Gong. Multi-labelled classification using maximum entropy method. In Proc. Of SIGIR , 2005.

Authors:

Shweta C. Dharmadhikari received her M.Tech(CSE) from BVDU, Pune ,Maharashtra, India. She is pursuing research in the area of Machine Learning from DAU,Indore,(M.P.),INDIA. Presently she is working as an Associate Professor in Department of Information Technology at Pune Institute of Computer Technology , Pune. Her areas of interest include Machine Learning, Text Mining and Information Technology.

Dr. Maya Ingle did her Ph.D in Computer Science from DAU, Indore (M.P.) INDIA, M.Tech (CS) from IIT, Kharagpur, INDIA, Post Graduate Diploma in Automatic Computation, University of Roorkee, INDIA, M.Sc. (Statistics) from DAU, Indore (M.P.) INDIA.She is presently working as Professor/ Senior System Analyst, School of Computer Science and Information Technology, DAU, Indore (M.P.) INDIA. She has over 100 research papers published in various International/ National Journals and Conferences. Her areas of interest include Software Engineering, Statistical Natural Language Processing, Usability Engineering, Agile computing, Natural Language Processing, Object Oriented Software Engineering.

Dr. Parag Kulkarni hold PhD from IIT Kharagpur. UGSM Monarch Business School - Switzerland conferred DSc - Higher Doctorate on him. He is the founder and Chief Scientist of EKLat Research where he has empowered businesses through machine learning, knowledge management, and systemic management. He has been working within the IT industry for over twenty years. The recipient of several awards, Dr. Kulkarni is a pioneer in the field of Systemic Machine Learning. He has over 120 research publications including more than half a dozen books and 3 patents. His areas of research and product development include M-maps, intelligent systems, text mining, image processing, decision systems, forecasting, IT strategy, artificial intelligence, and machine learning.