ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

A Comparative study of Classifiers' Performance for Gender Classification

Santanu Modak , Abhoy Chand Mondal

Department of Computer Science, University of Burdwan, Burdwan ,West Bengal ,India

ABSTRACT: -Reviewer gender classification is an important function of Sentiment Analysis system. Both supervised and unsupervised approach may be applied for gender classification. In this paper we used supervised machine learning approach. We use three different classifiers, namely Naïve Bayes Classifier, Maximum Entropy Classifier and Decision Tree Classifier respectively. We trained all classifiers using same training set and same feature function. Then we test the Accuracy, Precision, Recall, F1-measure of all test cases using same test set. Finally, we make an comparative study about performance of this classifiers.

KEYWORDS: naïve bayes classifier; maxent classifier; decision tree classifier; text classification; gender classification; classifier

I. INTRODUCTION

Classification problem can be defined in the following way-we have a set of classes, then we have to predict the class of given input object. Classification problem can be used beyond Information Retrieval like (1) image classification-to detect the image belongs to which class –landscape or portrait. (2) check e-mail which is spam or not.(3)sort the message coming from friend, family, office etc. classification can be done manually, but that is very time consuming, so we have to use computer for this purpose. We have to derive rules for each class.

Apart from manual classification and rule-based technique, we can use supervised machine learning for classification. For that approach, we need a training set. Classifiers will be trained using that training set. After that, we have to create a test set to check the accuracy and other measures of classification. Training set and Test set should be independent to each other. Each object of training set should be labeled manually, which is comparatively easy approach rather than derive rules. But feature selection is the main approach in this learning mechanism. Our goal in this classification is to detect the best class of a given document means high accuracy in Test data.

II. RELATED WORK

In [5], Moghaddam et al, use Support Vector Machine to classify gender from visual image with low resolution (21-by-12 pixels) processed from 1,755 images from the FERET face database. The performance of SVMs (3.4% error) is shown to be superior to traditional pattern classifiers (Linear, Quadratic, Fisher Linear Discriminant, Nearest-Neighbor) as well as more modern techniques such as Radial Basis Function (RBF) classifiers and large ensemble-RBF networks. In [6], Zehang, et al used Principal Component Analysis (PCA) to represent each image as a feature vector (i.e., eigen-features) in a low-dimensional space. Genetic Algorithms (GAs) are then employed to select a subset of features from the low-dimensional representation by disregarding certain eigenvectors that do not seem to encode important gender information. They used four different classifiers to test the accuracy, namely Bayes classifier, Neural Network (NN) classifier, Support Vector Machine (SVM) classifier, and a classifier based on Linear Discriminant Analysis (LDA). Out of them, error rate of SVM classifier was very low 4.7% from an average error rate of 8.9% using manually selected features. In [7], Malcolm, et al used an extended set of predominantly topic content-free e-mail document features such as style markers, structural characteristics and gender-preferential language features together with a support vector Machine Learning algorithm. In [8], Mukherjee et al propose two new technique to improve the current result. The first technique introduces a new class of features which are variable length POS sequence patterns mined from the training data using a sequence pattern mining algorithm. In [9], yan et al presented a Naıve Bayes classification approach to identify genders of weblog authors. In addition to features employed in traditional text categorization, they used weblog-specific features such as webpage background colors and emoticons. The second technique is a new feature selection method which is based on an ensemble of several feature selection criteria and approaches. In [10], Amasyalı et al used four different classifiers to detect 3 different areas such as determining the identification of a Turkish document's author, classifying documents according to text's genre and identifying a gender of an author. Naive Bayes, Support Vector Machine, C 4.5 and Random Forest were used as classification methods.

III. PROBLEM DEFINITION

In this paper, we use classifiers for Gender Classification, where input object is a name and classifier will predict it belongs to which class-male or female. Our set of class has only two members. To train classifier, we use name corpus, where 8000 different name are already present and each name are same. We use suffle() function to build train set and test set. So every time, training set and test set are different from earlier step. Then we derive feature_selection() function from name and train the classifier using that features. Finally, we check the Accuracy, Precision, Recall, F1measure of each classifier using Test Set and make a comparative study.

IV. FEATURE SELECTION

Feature Selection is the main criteria to train a classifier. After training, classifier test every object of test set using that feature. Generally, classifiers follow probabilistic model of Information Retrieval. So, classifier calculates probabilities of each class of input object and produce output class whose probability is heighest. Classifier trained using that feature from training set. For testing purpose, classifiers calculate the probability of each object using that feature and predict output.

V. CLASSIFIERS

1. NAÏVE-BAYES CLASSIFIER

Naïve-Bayes classifier is working based on Baye's theorem of conditional probability. When input become very high, this classifier should be used. This classifier builds using probabilistic model. Here only two class label are present. So, we used binary classification. Conditional probability is calculated with respect to every name in the test set. Like, if input name is X and class label is C, then it will calculate $P(X|C)$ and $P(X|{\sim}C)$ where $P(X|C)$ is probability of name X belongs to class label C and $P(X|{\sim}C)$ is probability of name X not belongs to class label C.

2. MAXIMUM ENTROPY CLASSIFIER

Maximum Entropy Classifier, also called Conditional Classifier, converts labelled feature sets to vector using encoding. The encoded vector is used to calculate weight of each feature that used to label the test data. Some parameters like "algorithm = iis", "trace", "max_iter", "min_lldelta" have been set to get more accurate results.
The basic idea behind Maximum Entropy Classifier is probabilistic distribution function. "iis" algorithm iteratively increases the weight. "max_iter" species maximum number of iterations where "min_lldelta" specifies least change in log_likelihood required for iteration and change the weights.

3.DECISION TREE CLASSIFIER

Decision Tree classifier works by creating classification tree, where each non-leaf node corresponds to a feature name and their children corresponds to a feature value. Decision Tree classifier is often used in text classification problem. This is also a supervised machine learning approach. So training set and test set need to be created. During training, Decision Tree Classifier creates a binary tree where the child nodes are also instance of classifier. The leaf nodes contain only a single label, which the intermediate child nodes contain decision mapping for each feature. It contributes the final decision of classification.

VI. SIMULATION RESULTS

Precision (P) is the fraction of retrieved documents that are relevant; P(relevant|retrieved)
Recall (R) is the fraction of relevant documents that are retrieved; P(retrieved|relevant).
If retrieved document is relevant, then it is true positives (tp), If retrieved document is not relevant, then it is false positives (fp), If not retrieved document is relevant, then it is false negatives (fn), If not retrieved document is not relevant, then it is true negatives (tn). So

$$Precision = tp/(tp+ fp)$$
$$Recall = tp/(tp+ fn).$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

Accuracy is the fraction of its classifications that are correct. So

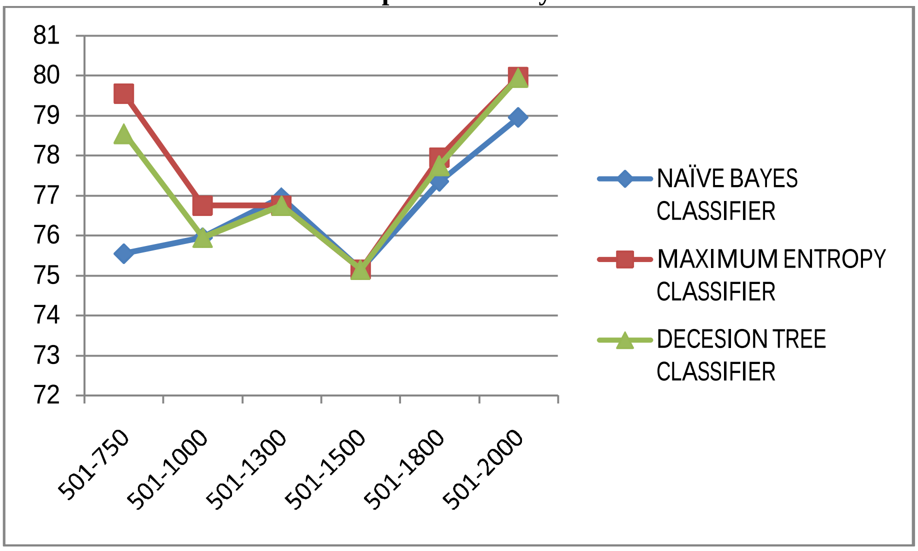
accuracy =(tp + tn)/(tp + f p + f n + tn).

A single measure that trades off precision versus recall is the *F measure*,which is the weighted harmonic mean of precision and recall.

Accuracy (comparative analysis)

TRAINING SET	NAIVE BAYES CLASSIFIER	MAXIMUM ENTROPY CLASSIFIER	DECESION TREE CLASSIFIER
501-750	0.755511022044	0.795591182365	0.785571142285
501-1000	0.759519038076	0.76753507014	0.759519038076
501-1300	0.769539078156	0.76753507014	0.76753507014
501-1500	0.751503006012	0.751503006012	0.751503006012
501-1800	0.773547094188	0.779559118236	0.77755511022
501-2000	0.789579158317	0.799599198397	0.799599198397

Graph based analysis



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

1) NAÏVE BAYES CLASSIFIER FOR MALE

TRAINING SET	Precision	Recall	F1 measure
501-750	0.714285714286	0.491329479769	0.582191780822
501-1000	0.696774193548	0.596685082873	0.642857142857
501-1300	0.677248677249	0.703296703297	0.690026954178
501-1500	0.638418079096	0.653179190751	0.645714285714
501-1800	0.698224852071	0.655555555556	0.676217765043
501-2000	0.728813559322	0.693548387097	0.710743801653

2) NAÏVE BAYES CLASSIFIER FOR FEMALE

TRAINING SET	Precision	Recall	F1 measure
501-750	0.768421052632	0.895705521472	0.827195467422
501-1000	0.787790697674	0.852201257862	0.818731117825
501-1300	0.825806451613	0.807570977918	0.81658692185
501-1500	0.813664596273	0.803680981595	0.808641975309
501-1800	0.812121212121	0.84012539185	0.825885978428
501-2000	0.82298136646	0.846645367412	0.834645669291

3) MAXIMUM ENTROPY CLASSIFIER FOR MALE

TRAINING SET	Precision	Recall	F1 measure
501-750	0.676616915423	0.78612716763	0.727272727273
501-1000	0.704402515723	0.618784530387	0.658823529412
501-1300	0.671875	0.708791208791	0.689839572193
501-1500	0.638418079096	0.653179190751	0.645714285714
501-1800	0.703488372093	0.672222222222	0.6875
501-2000	0.733695652174	0.725806451613	0.72972972973

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

4) MAXIMUM ENTROPY CLASSIFIER FOR FEMALE

TRAINING SET	Precision	Recall	F1 measure
501-750	0.875838926174	0.800613496933	0.836538461538
501-1000	0.797058823529	0.852201257862	0.823708206687
501-1300	0.827361563518	0.801261829653	0.814102564103
501-1500	0.813664596273	0.803680981595	0.808641975309
501-1800	0.819571865443	0.84012539185	0.829721362229
501-2000	0.838095238095	0.843450479233	0.84076433121

5) DECESION TREE CLASSIFIER FOR MALE

TRAINING SET	Precision	Recall	F1 measure
501-750	0.675531914894	0.734104046243	0.703601108033
501-1000	0.696774193548	0.596685082873	0.642857142857
501-1300	0.671875	0.708791208791	0.689839572193
501-1500	0.638418079096	0.653179190751	0.645714285714
501-1800	0.701754385965	0.666666666667	0.683760683761
501-2000	0.733695652174	0.725806451613	0.72972972973

6) DECESION TREE CLASSIFIER FOR FEMALE

TRAINING SET	Precision	Recall	F1 measure
501-750	0.852090032154	0.812883435583	0.832025117739
501-1000	0.787790697674	0.852201257862	0.818731117825
501-1300	0.827361563518	0.801261829653	0.814102564103
501-1500	0.813664596273	0.803680981595	0.808641975309
501-1800	0.817073170732	0.84012539185	0.828438948995
501-2000	0.838095238095	0.843450479233	0.84076433121

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

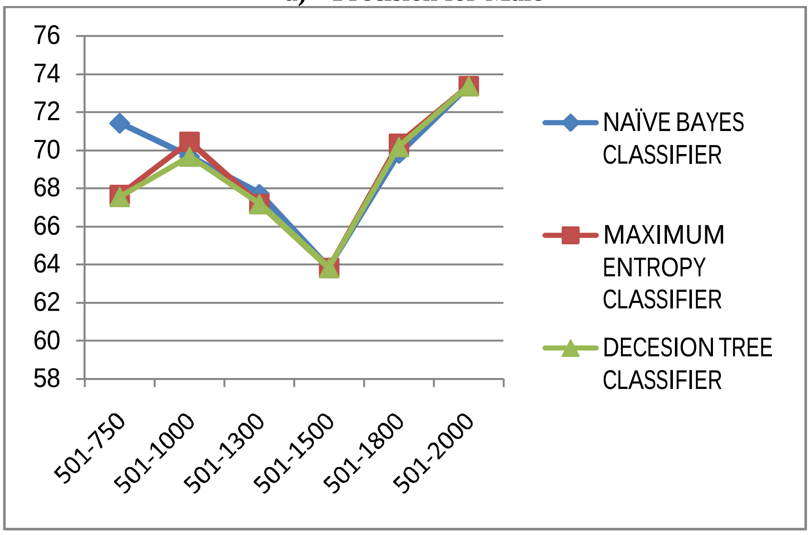
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

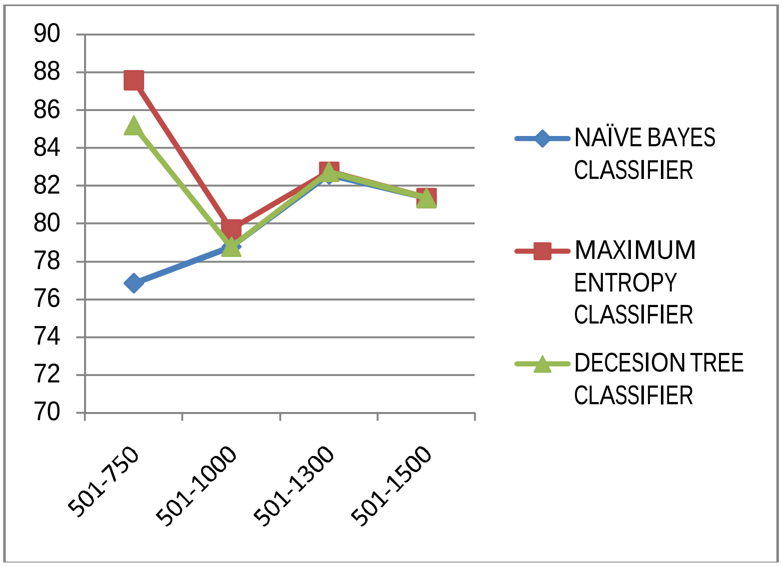
Vol. 2, Issue 5, May 2014

Graph Based Comparative Analysis

a) Precision for Male



b) Precision for Female



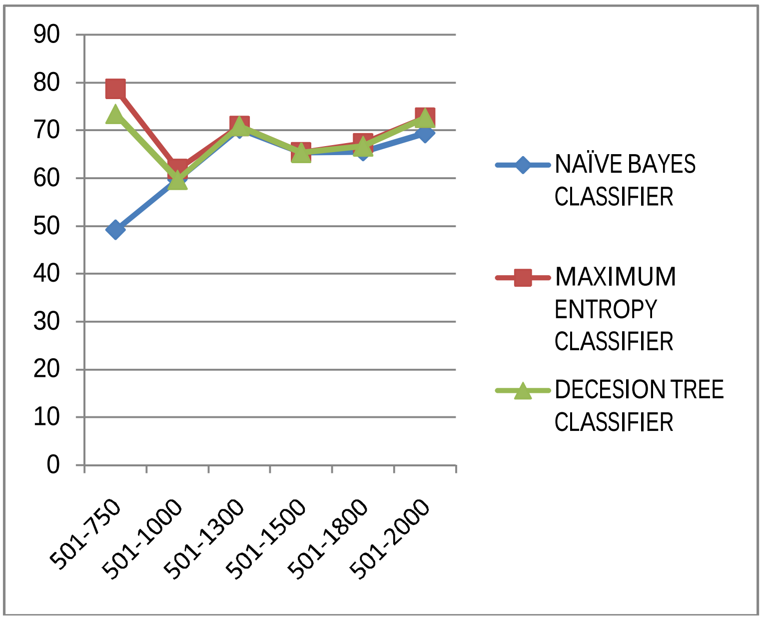
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

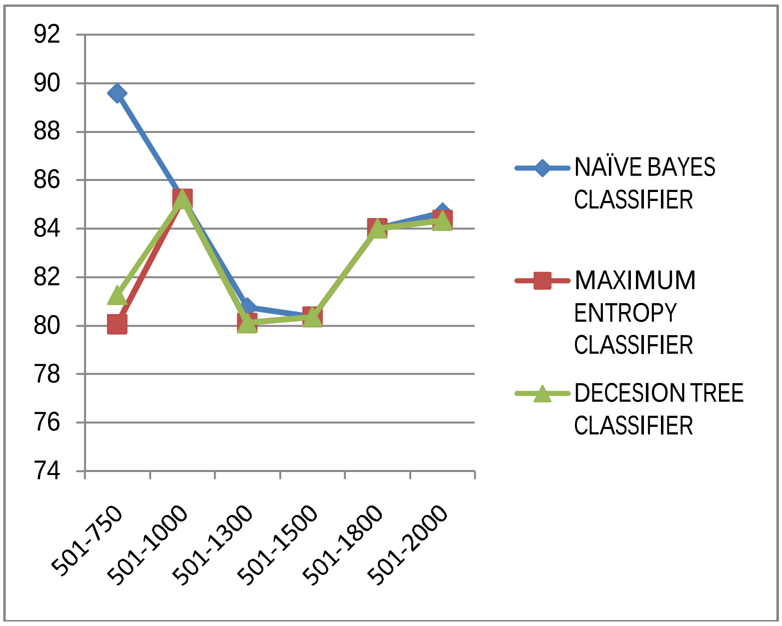
(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

c) Recall for Male



d) Recall for Female



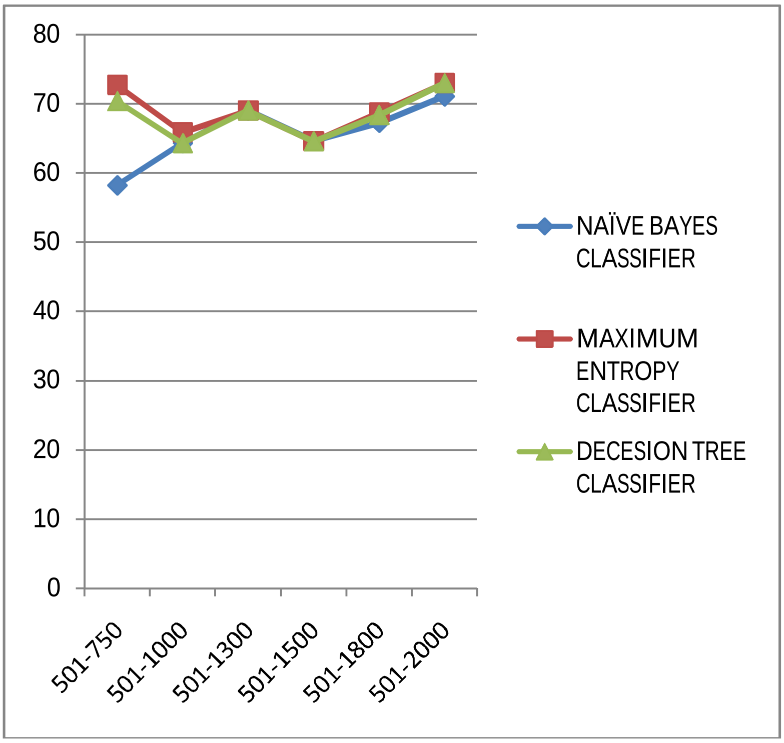
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

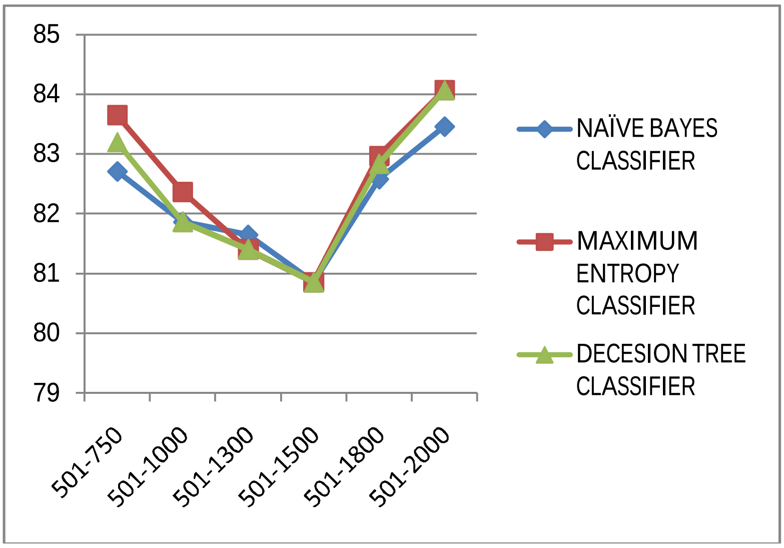
(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

e) F1 measure for Male



f) F1 measure for Female



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 5, May 2014

ACKNOWLEDGMENT

Natural Language Toolkit with Python 2.7 are used to get all results in this paper. Names Corpus, Version 1.3 (author: Mark Kantrowitz and Bill Rossis dated 1994-03-29) is used for training and testing purpose. All graphs are plotted by Microsoft Word chart.

VII. CONCLUSION AND FUTURE WORK

In this paper, we use three classifiers to check the accuracy of three classifiers. Based on the results, we conclude that Maximum Entropy Classifier with "iis" algorithm, gives best result compare to other classifier. In future, we will try to use this classifier for another type of text classification problem.

REFERENCES.

[1] Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.
[2] Willi Richert,Luis Pedro Coelho, Building Machine Learning Systems with Python, 2013 Packt Publishing
[3] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze(2009) An Introduction to Information Retrieval, Online edition, Cambridge UP
[4] Jacob Perkins, Python Text Processing with NLTK 2.0 Cookbook
[5] Moghaddam, Baback, and Ming-Hsuan Yang. "Gender classification with support vector machines." *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 2000.
[6] Sun, Zehang, et al. "Genetic feature subset selection for gender classification: A comparison study." *Applications of Computer Vision, 2002.(WACV 2002). Proceedings. Sixth IEEE Workshop on*. IEEE, 2002.
[7] Corney, Malcolm, et al. "Gender-preferential text mining of e-mail discourse." *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*. IEEE, 2002.
[8] Mukherjee, Arjun, and Bing Liu. "Improving gender classification of blog authors." *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*. Association for Computational Linguistics, 2010.
[9] Yan, Xiang, and Ling Yan. "Gender Classification of Weblog Authors." *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*. 2006.
Amasyalı, M. Fatih, and Banu Diri. "Automatic Turkish Text Categorization in terms of Author, genre and gender." *Natural Language Processing and Information Systems*. Springer Berlin Heidelberg, 2006. 221-226.

BIOGRAPHY

Santanu Modak is Junior Research Fellow in the Department of Computer Science, University of Burdwan. He received his B.Sc(Hons) and M.Sc in Computer Science degrees in 2009 and 2011 respectively. He qualified UGC-NET in Computer Science and Applications. He is a life member of Indian Science Congress Association (ISCA) and member of International Association of Computer Science and Information Technology (IACSIT).

Abhoy Chand Mondal is currently Associate Professor and Head of Department of Computer Science, Burdwan University, W.B., India. He received his B.Sc.(Mathematics Hons.) from The University of Burdwan in 1987, M.Sc. (Math) and M.C.A. from Jadavpur University, in 1989, 1992 respectively. He received his Ph.D. from Burdwan University in 2004. He has 1 year industry experience and 18 years of teaching and research experience. No. of papers more than 50 and no of journal is 25.