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. 4, Issue 6, June 2016

Sentimental Analysis of Movie Review using Machine Learning Algorithm with Tuned Hyperparameter

Suchita V. Wawre[1], Sachin N. Deshmukh[2]

M.Tech Student, Dept. of Computer Science and Information Technology, Dr B. A. M. University Aurangabad, Maharashtra, India

Professor, Dept. of Computer Science and Information Technology, Dr B. A. M. University Aurangabad, Maharashtra, India

ABSTRACT: Today Internet is reach source of information as large amount of information is available online. Discussion forum, review sites, blogs are some of the rich resources where review are available. Sentiment analysis system classify these review into positive, negative or neutral class which further can be use by customer to make choice of product and by businessmen for finding customer satisfaction and to make business policies. The aim of the paper is to classify movie review in to positive or negative polarity by using machine learning algorithm such as Naïve bayes, support vector machine, and random forest. This approach make the use of machine learning classifier's set of configuration parameters known as hyperparameters of random forest and svm, which are required to be tuned before a model gets trained. In this approach if classifier make use of this hyperparameter then random forest and svm model result into high accuracy. The results obtain by this approach provide accuracy of 84.29% for naive bayes, 96% for svm, 95% for random forest. After making use of hyperparameter the accuracy increased with 97.42% for svm, and 96% for random forest.

KEYWORDS: Sentimental Analysis, supervised Algorithm, Naive bayes, Support vector machine**,** Hyperparameters.

I. INTRODUCTION

Sentiment analysis focuses on sentiment which is express as positive or negative sentiments. Sentiment Analysis is a process classifying user review and determining the sentiment of users towards subject matter such as product, person. Sentiment can be Positive, Negative or may be Neutral Sentiment. Opinions are very important in decision making because whenever anybody needs to make a decision, he or she want to know others' opinions. For business businessmen and organizations want to know consumer opinions about their products and services. Opinion business men to make business policies and strategies .For Individual customer also want to know the opinions of people about product before buying that product, and others' opinions before making a voting decision about political candidates in a political election. In the past, an individual were making decision by asking opinion of their friends and family. An organization or a business conduct surveys, opinion polls when they need public or customer opinions about product. Acquiring public and consumer opinions is huge business itself for marketing, public relations, and political campaign companies. Nowadays, if customer wants to buy product, one is no longer limited to asking friends and family for opinions because increasing use internet reviews site and discussions forum on the Web about the product. For organization there is no need to conduct surveys, opinion polls in order to gather public opinions because there is large information publicly available online.

A challenging aspect of sentimental analysis by using machine learning is that they are not identifiable by keyword. For example "how could anyone watch this song" contain no single word which convey positive or negative sentiment. Consider another sentence "story of movie is good but visually its boring to watch. This sentence contains both positive and negative sentiment. In sentence "story is good" is positive sentiment while "visually its boring to watch" is negative sentence.

DOI: 10.15680/IJIRCCE.2016. 0406303

Another challenging aspect is that the traditional text processing considers that a little change in two bits of content has no change in the significance or meaning [1]. But in sentiment analysis a little change in two bits of content has change in the significance or meaning, consider Example “story is good” is different from “the story is not good”. In sentiment analysis an opinion word which is considered as a positive in one situation may be considered as negative in another situation. Twitter contains more informal text which user can understand but system cannot. For example “the movie is same as its last movie”. In this success of movie depend on last movie.

A. Levels of Sentiment Analysis

Sentiment analysis can mainly be classified into three level sentence level, document level, aspect level.

- Document level: Document level sentimental analysis is analysis to determine whether a whole document is positive or negative sentiment. For example, given a movie review, the system determines polarity of movie review whether the review is positive or negative. This is known as document-level sentiment classification. Document level focus on single entity and it is not applicable on to documents which compare multiple entities.

- Sentence level: Sentences level sentimental analysis is analysis to determine whether each sentence expressed a positive, negative, or neutral opinion. For example given tweet system can be classifies as positive, negative or neutral.

- Aspect level: Classifies sentences/documents as positive, negative or neutral based on the aspects of the sentences or documents commonly known as aspect-level sentiment classification. It is based on the idea that an opinion. It consists of a sentiment (positive or negative) and a target (of opinion). For example “The iPhone’s sound quality is good, but its battery life is short” evaluates two aspects, sound quality and battery life, of iPhone. The sentiment on iPhone’s sound quality is positive, but the sentiment on its battery life is negative. The call quality and battery life of iPhone are the opinion targets.

II. Related work

The sentiment analysis and opinion mining were first introduced in the year 2003.

In paper[1] make the use of movie reviews as dataset, the paper make use of standard machine learning techniques such as Naive Bayes, maximum entropy classification, and support vector machines which out-perform human produced baselines. But the three machine learning methods we employed do not perform as good as sentiment classification as on traditional topic-based categorization. In Performance Naive Bayes perform the worst and SVMs perform best.

In paper[2] is based on Document level opinion mining system that classify the documents as positive, negative and neutral. Proposed system also handled Negation. Experimental make use of movies dataset from IMDB. It make the use of POS tagging. Result show that Document based Sentiment Orientation System performs well with respect to the movie domain as compared to ‘AIRC Sentiment Analyzer.

In paper[3] compared three supervised machine learning algorithms such as SVM, Naive Bayes and kNN for sentiment classification of the movie reviews dataset that contains 1000 positive and 1000 negative processed reviews. The approch show that the SVM approach outperformed than the Naïve Bayies and k-NN approaches and the training dataset had a large number of reviews. The SVM approach acquires accuracies of more than 80% in classifying data correctly. When large training dataset containing 800 to 1000 reviews were used will perform better in sentiment classification for all three algorithms for the reviews about movie reviews.

In paper[4] proposes a sentiment analysis of movie reviews using a natural language processing and machine learning approaches in combination . Data pre-processing was done on the datasetfirstly . Secondly, the two classifiers, Naive Bayes and SVM, is used in combination with different feature selection schemes to obtain the results for sentiment analysis. Thirdly, the model for sentiment analysis is extended to obtain the results for higher order n-grams. The classification of movie review results show that Linear SVM classifier gives more accuracy than Naïve Bayes classifier. The results obtained for linear SVM are also better than the previous works. The accuracy increases for the bigrams . The affect of varying different parameters is different is shown in this work.

In paper[5] effective way to perform distant supervised learning is by using emoticons for training data. This approach of classifying sentiment different machine learning algorithms such as Naïve Bayes, maximum entropy classification can achieve high accuracy. Although Twitter messages have unique characteristics, machine learning algorithms classify tweet sentiment with similar performance.

In paper [6] propose make use of new approach called as Combined Approach to perform sentiment analysis on movie review. This approach combines two separate classifier Support Vector Machine (SVM) and Hidden Markov Model (HMM). By making use of classifier combine rule it combines results of these classifier. With the use of two classifier and classifier together by using combination rules it is possible to improve classification results. Classifier handles slang words and smiley. This proposed approach result in good sentiment classification with higher accuracy.

In paper [7] This paper make use supervised learning technique called the Random forests for classification of data by changing the values of different hyper parameters in Random Forests Classifier to get high accuracy of classification results. This paper also focuses on experimental comparison of Random Forests classifier with some supervised learning technique like NB (Naïve Bayes), C4.5 and ID3 with respect to their accuracy of correctly classified instances, incorrectly classified instances and very important ROC Area**.** Result analysis shows that Random Forests outperforms all the three classifiers *NB (Naive Bayes), C4.5 and ID3* in terms of correctly and incorrectly classified instances and ROC Area.

In paper[8] focus on the relevancy of features is an important property that should be taken into account for determining a parametrization rule for the Forest-RI algorithm. This paper shows Influence of Hyperparameters on Random Forest Accuracy.

In paper [9] focused on using the Random Forest to perform sentiments analysis of movie review dataset by Tuning of hyperparameters in random forest. On the basis of experimental results, random forest performed well on the movie review datasets. For dataset V1.0 it result in high accuracy of 87.85% and for dataset V2.0 it provided results with accuracy of 91.00%. Most of the previous work has focused on Support Vector Machine, Naive Bayes and Maximum Entropy for the sentiment classification, but according to experiments carried out in this paper random forest can provide better results if hyperparameters are fine tuned.

III.METHODOLOGY

A. Data Collection

This paper uses the Internet Movies Database (IMDB) movie review dataset. It makes use of standard dataset available at http://www.cs.cornell.edu/people/pabo/movie-review-data/. Dataset is Movie Review Dataset V1.0 which consist of 1400 movie review out of which 700 reviews are positive and 700 reviews are negative.

B. Text Preprocessing

This stage includes getting the actual text of review and each review in single line. As a result, this method will turn into just splitting the content of the file by the end of the line character. To get matches with the AFINN data that we used we convert the reviews into lower case. Also omitted punctuations, numbers and control characters to get correct matches.

C. Classification Algorithms

Let f1, f2,.. , fm be a predefined set of m features that can appear in a document. Let ni(d) denote number of times features fi occurs in document d. Then, each document d can be represented by the document vector
d := (n1(d), n2(d),…. , nm(d)).

I. Naive bayes

Naive bayes classifier is based on the Bayesian probability. The Naive Bayes classifier assume that probabilities of features are independent of one another that means one feature in the document is independent of other feature. Document is collection of words and assumes that the probability of a word in the document is independent word and its position of word in the document .We derive the Naive Bayes (NB) classifier by Bayes' rule, in eq. (1)

$$p(c/d)=\frac{p(c)p(d/c)}{p(d)} \quad \text{eq. (1)}$$

Where P(d) plays no role in selecting c. But its conditional independence assumption clearly does not hold in real-world.

II. Support vector machines

The document can be efficiently classified by using svm classification algorithm. The basic idea behind SVM classification is to separates the document vector in one class from the other with maximum margin by finding the hyper-plane with maximum margin. That why this classifier are called large margin classifier. They are not probabilistic classifiers like Naïve Bayes. This search corresponds to a constrained optimization problem; let the class cj $\{1, -1\}$ be the correct class of document denoted by dj, the solution can be given by vector W in eq. (2)

$$w := \sum_j \propto jcjdj, \propto j \geq 0 \quad \text{eq. (2)}$$

Where the αj 's can be obtained by solving a problem of dual optimization. Those document dj such that αj is greater than zero are called support vectors, because αj are the only document vectors contributing to vector w. Classification of instances consists of finding which side of w's hyper plane they fall on

III. Random Forest

Random forests implement decision tree algorithms to learn a number of decision trees to make a classification decision. At every step of the learning process an attribute is selected to split to two or more different parts and this process is iteratively repeated until the attributes are exhausted or we have reached a pure classification split. A pure classification split is when the split parts represent only one class that they belong to. At every split we try to reach a local optimum solution.

iv. Hyperparameters of Random Forest

As Random Forest is the combination of decision Trees, it deals with multiple number of hyperparameters which are:

- Number of Trees to construct for the Decision Forest
- Number of features to select at random
- Depth of each trees.

All hyperparameters are set manually such a way that it gives good results for the parameter that we have set manually. Each hyperparameters effect the output prediction. First hyperparameter is Number of Trees in the forest. From experiments its is seen that increase in number of trees increase accuracy of the random model. Larger the forest size better the accuracy, but the accuracy will remain stable at certain point even there is an increase in number of trees. Second parameter is Number of features which play important role in classification. Random forest have two values of features which are very famous 13 and 14 and may provide good accuracy results compared to other values of features. Random forest did not take all features. Depth of tree is also a very important hyperparameter in random forest. Model will suffer from under fitting if smaller value is chosen. Influence of these hyperparameters is discussed in Experiments and Results.

v. Hyperparameters of svm

Support vector machine deals with multiple number of hyperparameters which are:

- C cost
- sigma

c and sigma two hyperparameter are used and set such a way that it gives good result.Each parameter as effect on prediction. First hyperparameter is c cost parameter of svm, from experiments its is seen that increase in c increase accuracy of the svm model. Larger the value of c better the accuracy, but the accuracy will remain stabble at certain point even there is an increase in value of c. Second parameter is sigma correspond to kernel used to map input data into feature space. From experiment it was found that increase in sigma decrease the accuracy.

IV.EXPERIMENT

All the three algorithms are applies on the available sentiment analysis standard datasets. Standard movie review datasets V1.0 movie review dataset of Cornell University are used. All preprocessing is done on the dataset v1.0 movie review dataset. After preprocessing all algorithms are performed. Algorithm makes the use of cross validation function. Cross validation is a good method to evaluate the performance of a model as it divides the data into two parts. One part is used to train the model and build the classifier and the second part is used to test the accuracy of the predictions of the model. This process is done iteratively and different subsets of the data are used for training and testing at each iteration

iv.i. Naive Bayes

Below Table1 is the confusion matrix for the naive bayes classifier obtained after classification. The classifier has obtained accuracy of 84.29 % for dataset v1.0 sentence polarity dataset.

Table 1**:** Confusion Matrix of Naive bayes for v1.0 movie review dataset

	Actual		
predicted		Positive	Negative
	positive	26	5
	Negative	17	92

iv.ii. Support Vector Machine

Below Table2 is the confusion matrix of the performance of the support vector machine. We can see that this classifier has more accuracy as compared to naïve bayes. The model obtained the accuracy of 95.71 % for dataset. After the use of hyperparameter c with value 4 and sigma 0.25 the accuracy increase to 97.42%

Table 2**:** Confusion Matrix of support vector machine for v1.0 movie review dataset

	Actual		
predicted		Positive	Negative
	positive	41	4
	Negative	2	93

iv.iii. Random Forest

Below Table3 is the confusion matrix of the performance of the Random forest. The model obtained the accuracy of 94.45% for dataset. After the use of hyperparameter ntree with value 600 and mtry value 2 the accuracy increases to 96%.

Table 3: Confusion Matrix of Random Forest for v1.0 movie review dataset

	Actual		
predicted		Positive	Negative
	positive	40	4
	Negative	3	93

V. EFFECT OF HYPERPARAMETER ON ACCURACY

v.i. Effect of support vector machine hyperparameter on accuracy

Below mentioned is accuracy results of Support vector machine for dataset V1.0 movie review dataset that are achieved by automatically changing values of all different hyperparameters. Svm make use of sigma and c hyperparameter. For Support vector machine different values of c and sigma different accuracy is obtained. From experiment it was seen that increase in value of c increases the accuracy and increase in sigma value of kernel decreases value of accuracy of model. This is shown by graph below. Before use of hyperparameter the model obtained the accuracy of 95.71 % for dataset. After the use of hyperparameter c with value 4 and sigma 0.25 the accuracy increases to 97.42%. The different values of c and sigma different accuracy is obtain can be shown below. Figure 1 describe accuracy obtain at different values of C and Sigma.

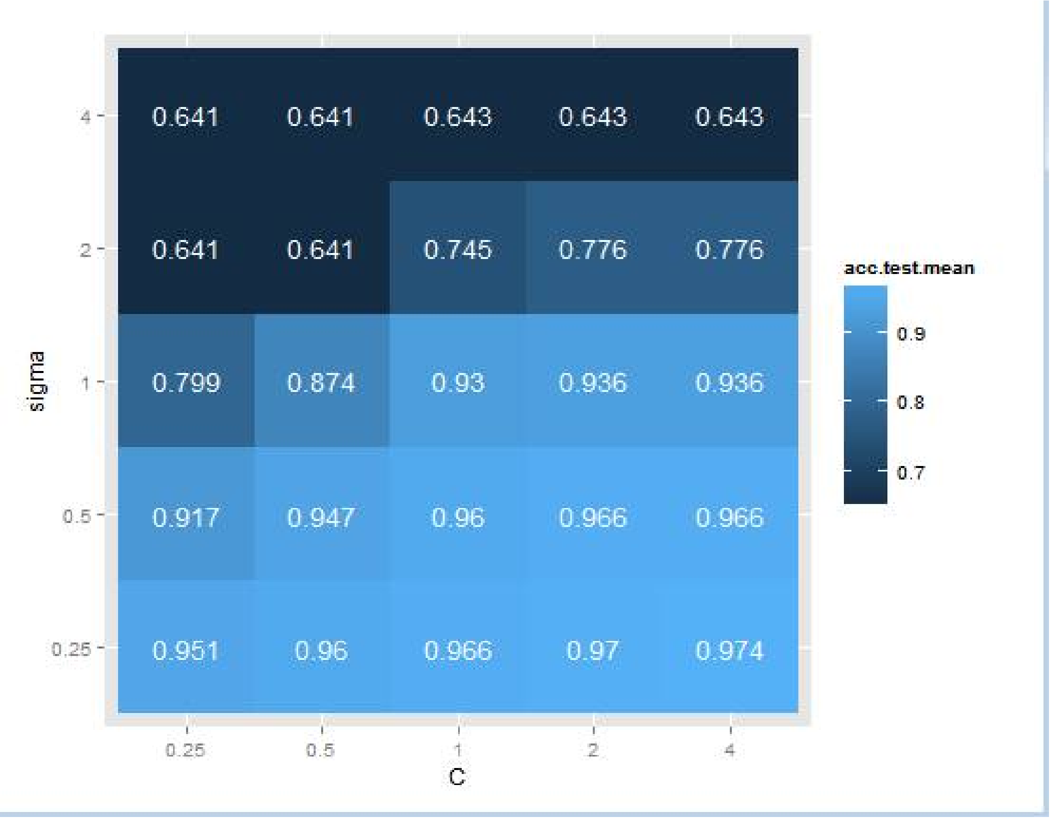

Figure 1: Influence of hyperparameter sigma and c on accuracy

v.ii Effect of Random forest hyperparameter on accuracy

Different number of hyperparameter are tried for the datasets to obtain good result. We have considered two hyperparameter that is number of trees and number of features. Proposed approach we found that if the number of trees increases then accuracy will increase up to certain values and after that it will be stable. For number of feature 2 it provide high accuracy. Before use of hyperparameter the random forest provides the accuracy of 94.45%. After setting values of random forest hyperparameter number of tress to 600 and number of features to 2 the model obtain the accuracy of 95.16%. For different values of ntree and mtry different accuracy is obtain and shown below. Figure 2 describe accuracy obtain at different values of mtry and ntree.

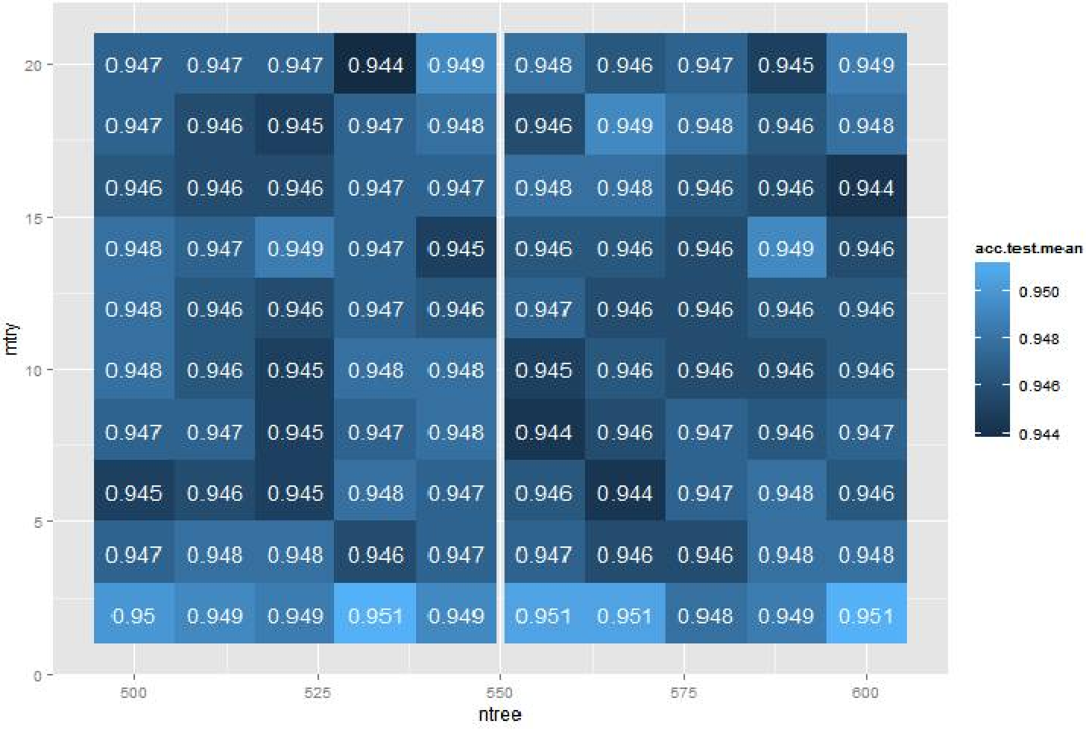

Figure 2: Influence of hyperparameter mtry and ntree on accuracy

VI. Result Analysis

Different machine learning algorithms are applies on sentiment analysis standard datasets. The standard movie review datasets V1.0 containing 1400 review from Cornell University [11] is selected, which is been used by many of the researchers in the field of sentiment analysis. After applying naïve bayes, svm, random forest machine learning algorithm on datasets, results achieved are very much promising and competitive. The comparative results of different researchers can be shown in table 4 below.

Table4. Experimental result on dataset v1.0

Sr.No	Author	Approach	Accuracy
1.	Pang and Lee [1]	Naïve Bayes, SVM, Maximum Entropy	82.90%
2.	Mullen and Collier [10]	Support Vector Machine	86.00%
3.	Hitesh Pawar, Sanjay Banderi, Glory shah [9]	Random Forest	87.85%
4.	Proposed Approach	Naïve Bayes, SVM, Random Forest	97.42%

VII. Conclusion

This paper makes the of machine learning algorithm such as naïve bayes, svm, and Random Forest to do sentiments analysis of movie review dataset. Different machine learning model obtain different accuracy. The propose approach

Support vector machine model as high accuracy as compare to naïve bayes and random forest. The accuracy of svm and random forest model can further increased by use of hyperparameter. Table 5. describe different machine learning model and its accuracies.

Table 5**:** Result accuracy for v1.0 sentence polarity dataset

Method	Accuracy
Naive Bayes	84.29%
Support Vector Machine	95.71%
Support Vector Machine after tuning of hyperparameter	97.42%
Random Forest	94.45%
Random forest after tuning of hyperparameter	95.14%

VIII. FUTURE WORK

We will make feature selection by using unigrams, bigrams and trigrams feature in order to increase the accuracy of model

REFERENCES

1. Bo Pang and Lillian Lee and Shivakumar Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques”, Language Processing (EMNLP), Philadelphia, pp. 79-86, July 2002.
2. Richa Sharma, Shweta Nigam and Rekha Jain, “Opinion mining of movie review at document level”, International Journal on Information Theory (IJIT), Vol.3,No.3, July 2014.
3. P.Kalaivani, Dr.K.L.Shunmuganathan, “Sentiment classification of movie review by supervise machine learning approach”, Indian Journal of Computer Science and Engineering (IJCSE) ,Vol. 4 No.4, Aug-Sep 2013
4. Gautami Tripathi and Naganna S, “Feature Selection and classification approcha for Sentiment Analysis”, Machine Learning and Applications: An International Journal (MLAIJ) ,Vol.2, No.2, June 2015
5. Hemalatha1, Dr. G. P Saradhi Varma, Dr. A.Govardhan,“Sentiment Analysis Tool using Machine Learning Algorithms”,International Journal of Emerging Trends & Technology in Computer Science Volume 2, Issue 2, March – April 2013
6. Anurag Mulkalwar, Kavita Kelkar “Sentiment Analysis on Movie Reviews Based on Combined Approach”, International Journal of Science and Research, Volume 3 Issue 7, July 2014
7. Mr. Hitesh H. Parmar, Prof. Glory H. Shah, “Experimental and comparative analysis of machine Learning classifier” , International journal of Advance research in computer and software engineering, volume 3, Issue 10,oct 2013
8. Simon Bernard, Laurent Heutte, and Sebastie, “Influence of hyperparameters on random Forest Accuracy”, MCS, Springer, pp.171-180, 2009, Lecture Notes in Computer Science, vol. 5519
9. Mr. Hitesh H. Parmar, Prof. Glory H. Shah, Sanjay Bhanderi “Sentiment Mining of Movie Reviews using Random Forest with Tuned Hyperparameters”
10. T. Mullen, N. Collier, “Incorporating topic information into sentiment analysis models”, In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, Article 25
11. Dataset: cs.cornell.edu/people/pabo/moviereview-data/.

BIOGRAPHY

Suchita V. Wawre Received B.E in Computer science and Engineering from MSS’s College Of Engineering and Technology, nagewadi, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad.. Currently pursuing M.Tech in Computer Science and Engineering from Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India.

Dr. Sachin N. Deshmukh completed his Ph.D. and M.E in Computer Science and Engineering. He is currently a Professor in Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India. His research area is “Text mining, Social Web mining and Intension Mining”. He is a member of Adhoc Board of Studies in BioInformatics and Liberal arts of Dr. B. A. M. University Aurangabad and Adhoc Board of Computer Science at Shivaji University Kolhapur.