A Consumer Behavior Prediction Method for E-Commerce Application

Kareena, Raj Kumar

Abstract*: The consumer behavior analysis is the technique which is applied to analyze consumer behavior. The customer behavior analysis has the three steps which are pre-processing, feature extraction and classification for prediction. In the previous work, Naïve Bayes was applied for the consumer behavior analysis. In this work, hybrid classifier is designed for the customer behavior analysis using Decision Tree and KNN. The proposed method is implemented in anaconda python and results are compared with the previously used Naïve Bayes method, for this analysis consumer reviews from Amazon website are used.

Index Terms*: Consumer behavior, Decision Tree, KNN, Naïve Bayes.

I. INTRODUCTION TO CONSUMER BEHAVIOR

Over the years, e-commerce has gained huge popularity. Due to its extensive comfort and time-saving properties people prefer online shopping today. There are several retail websites online which provide a platform for clients to buy millions of products online. Amazon and Snapdeal are the most commonly known online websites where customers can also sell and purchase products online [1]. It is also possible for the clients to post reviews related to certain products online so that the other consumers can have a viewpoint regarding the products which they want to purchase. Highly valuable information related to the qualities of products is provided through the customer reviews [2]. However, these reviews are very confusing and disorganized due to which it is difficult for an individual user to summarize such large numbers of reviews at one time. Thus, it is important to provide an easy mechanism through which the reviews of users can be analyzed and evaluated. Aspects play a very important role in reviews. An attribute of a particular product is called an aspect and one product possibly has several aspects [3]. Certain aspects can be more important than the others in terms of the impacts caused by them on decision making process of consumers as well as the organizations and product development strategies. For instance, aspects like "picture quality" and "lenses" have higher impact in comparison to the "wrist strap" aspect in case of a Nikon camera product [4]. Not only the quality of products but also their reputation can be improved through the reviews of important aspects of the product as per the organizations. The manual identification of these aspects from the reviews of the product is very difficult and thus, various approaches are needed for this [5]. The word-to-mouth reference or recommendation of a product will benefit the producer if a consumer who is highly satisfied by a product. The costs and efforts of others can be reduced alternatively for other individuals as well. However, a consumer can have a strong sense of justice of informing others about the quality of a product whether he/she is satisfied or not with the product [6]. The other individuals can be prevented from facing similar experiences with the help of that reviews. Irrespective of any of these situations, the real feelings and actual detail description provided by genuine users helps other customers collect lots of information about the quality of a product [7]. The consumers can thus make better purchase decisions with the help of these reviews.

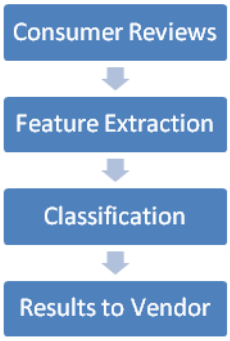

Fig. 1. General Process of Consumer Behavior Analysis [23]

Fig. 1. Shows the general steps that are involved when analyzing the customer's behavior on the basis of reviews collected related to the products. In the first step the reviews of customers are collected from different sources. The important data related to the reviews is extracted and separated from the remaining extra data posted by the customer in the second step. After the extraction of specific reviews, the important features from the review are extracted [8]. These extracted features define the polarity of the review such as the review is "positive", "negative" or "neutral". After the extraction of features, they need to be classified through feature classification step [9]. The features are classified into different categories based on their

Revised Manuscript Received on September 10, 2019.

Kareena, University Institute of Engineering (CSE), Chandigarh University, Gharuan, Mohali, Punjab, India.
Er. Raj Kumar, University Institute of Engineering (CSE), Chandigarh University, Gharuan, Mohali, Punjab, India.

respective properties and polarities. In order to perform feature classification, particular classifier Decision Tree is used among several existing classifiers like SVM, KNN, Naïve Bayes, and so on [10].Then KNN is used for final classification and the results are finally used by vendor for further action for business purposes. It means that the proposed algorithm is the hybrid of Decision Tree and KNN [18].

II. LITERATURE REVIEW

Yuanlin Chen, et.al (2015) studied the major actors that were in relevance to the helpfulness of reviews from the viewpoint of a client [11]. Initially, the three hypotheses related to review helpfulness were proposed on the basis of explanatory theory. The behavioral and psychology theory were included within this study. The major factors that affected the helpfulness of reviews from various dimensions were discussed from three hypotheses. The time and efforts required by a client to identify the required reviews of a particular product were saved by the information achieved through the proposed approach.

Hernandez Sergio, et.al (2017) proposed a linear-temporal logic model checking mechanism through which the structured e-commerce web logs were analyzed [12]. It was possible to convert the web logs into event logs which captured the behaviors of users by defining a general method of mapping log records based on the e-commerce structure. Further, for identifying the various behavioral patterns which included various actions performed by a client in a session, several predefined queries could be performed here. Towards the end, a real case study of a Spanish e-commerce website was used to test the efficiency of proposed approach. It was seen that the proposed approach provided highly effective outcomes.

Ru Jia, et.al (2017) proposed a novel classification approach through which the purchasing strategies of clients could be predicted [13]. The previously existing approaches included data rating which was not required in this proposed approach which included Bayesian classification mechanism. For performing classification, the user clicking behavior features were applied using Naïve Bayes classifier since it was commonly known for its simplicity and high efficiency. The outcomes proved the effectiveness of proposed approach as compared to previous approaches.

Ting Bai, et.al (2018) presented a study that focused on analyzing the behavioral factors of early reviewers on the basis of the reviews provided by them online [14]. Two commonly used e-commerce platforms named Amazon and Yelp were used in this research. Early, majority and laggards are the three consecutive steps in which the product lifetime was categorized. It was seen that a higher average rating score was assigned to an early reviewer. Also, more helpful reviews were posted by an early reviewer. It was also seen that the popularity of product was influenced by early reviewers' rating and their received helpfulness scores.

Namuk Ko, et.al (2018) proposed a technique for identifying the product opportunities from reviews of clients posted on social media [15]. Based on the several review posts posted by the clients for a given product, a topic modeling was designed by this research. On the basis of co-occurrences related to the topics present in every post, a keygraph was constructed here. Further, for generating new product opportunities from chance nodes achieved from keygraph, the chance discovery theory was applied. On the basis of large-scale and real-time VOC, the systematic ideation process was contributed for product opportunity analysis in this research.

Yusheng Zhou et.al (2019) proposed a random forest mechanism which used numerical and textual properties for predicting the review helpfulness [16]. It was seen through this research that for predicting the helpfulness of online reviews, the most important factor was review length. It was also seen that based on three various review types the importance of numerical properties was higher in comparison to the textual properties. Therefore, it was seen that to differentiate a helpful product review, the reviewed reputation was considered as an important predictor.

III. CLASSIFIER USED FOR HYBRID CLASSIFICATION

This research work is related customer behavior analysis. The customer behavior analysis technique which is designed in this research work will be based on the hybrid classification. The hybrid classification approach will be the combination of two classifiers which are Meta classifiers and base classifier. The base classifier will extract the features of the dataset means it will establish relation between target set and attribute. In the last, the Meta classifier will be applied which can generate the final classified result of prediction. In this work, the base classifier will be decision tree and Meta classifier will be KNN. The explanation is given below:-

Base-classifier: The classifier that is used within its default parametric settings is called the base classifier. Minimum two components named the predicted class and class probability distribution are included within the output of every base-level classifier for example the test set. A supervised learning algorithm which is mostly categorized into classification problems is called decision tree. The population or sample is categorized into two or more homogeneous sets on the basis of most significant splitter in the input variables in this technique. A tree structure is generated through decision tree by generating classification or regression models. A dataset is broken down into smaller subsets while at similar time duration an associated decision tree is designed incrementally. There are decision nodes and leaf nodes generated in the final result. There are two or more branches included in the decision node. A classification or decision is represented through a leaf node. Root node is the topmost decision node where a tree corresponds to the best predictor. Both categorical and numerical data are handled through decision trees. Depending upon the type of target variable, the types of decision tree are categorized. The decision tree that includes

categorical target variable is included in the categorical variable decision tree. If the decision tree includes continuous target variable it is categorized as continuous variable decision tree.

Meta-classifier: When the classifiers are combined, meta-classification is performed. The meta-classification includes three important steps. Multiple training subsets are generated from a training set in the initial step. Further, based on the algorithm and data training subset, every classifier is constructed individually in the second step. The results of base classifiers are integrated in the third step. The final higher-level step is called the Meta-classifier in which the final outcomes are achieved. The samples are classified by this classifier on the basis of nearest training samples. In the training process, the feature vectors as well as the labels of training images are stored. During the classification process, the unlabelled question point is ruled out when labeling the k-nearest neighbors. The characterization of object is done through the majority share based on the labels of k-NN. In the event where k=1 the object is classified on priority basis since that object is nearest to it. In the presence of only two classes, k is known to be an odd integer. The possibility of a tie to occur also arises when k is an odd whole number at the time of performance of multiclass categorization. The most important task of KNN classifier is to classify the samples on the basis of majority class of its nearest neighbor.

$$Class= \llbracket arg \rrbracket \_v \ max\sum\_((X\_i,y\_i)\in D\_z)\llbracket I(v=y\_i) \rrbracket \quad \ldots (1)$$

In the above equation (1), v represents the class label and y_i represents the class label for ith nearest neighbors, i denote an indicator function where value 1 is returned if the argument is true or else 0 values is returned. Therefore, the assignment of samples is done in the class of k-NN. It is important to select an appropriate similarity function and value of parameter k for making the recognition task successful.

Data Collection: Initially the data is collected of consumer reviews of Amazon e-commerce website from UCI repository which contains the reviews regarding various products of Amazon having varying polarity like some are positive, some are negative and some are neutral.

Pre-processor: After collection of data from UCI repository, data is preprocessed. It contains three steps that are:

Tokenizing: It means that dividing the paragraph into a different set of statements or it can be dividing a statement into different [17] set of words. For example, **The movie was great!** Then this statement will be tokenized as

The	movie	was	great	!

Cleaning: It means removing the special characters. For example we have exclamatory sign in the above sentence then this sign will be removed, the remaining part of sentence will be

The	movie	was	great

Stopword Removal: In this process all the words that do not add any value to the analysis is removed like is, am, are, was, he, she etc.

IV. RESEARCH METHODOLOGY

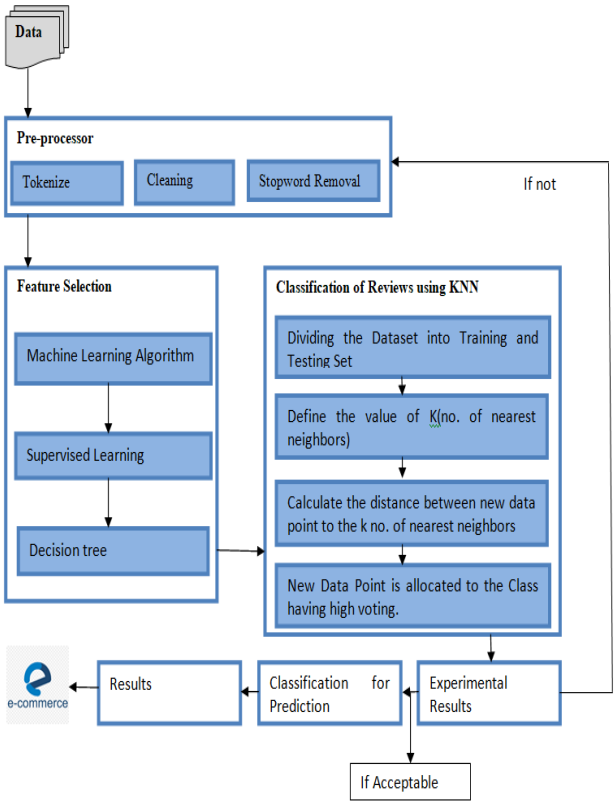

Fig. 2. RESEARCH METHODOLOGY

So the remaining part of sentence for analysis will be

movie	great

Feature Selection: In this process number of attributes is minimized into better subsets so that the accuracy could be more, the problem of overfitting could be removed and the training time could be minimized [19]. For this process machine learning algorithm is used called Decision Tree which is supervised learning algorithm. Each sentence is analyzed and sentiment score is given [22].

Classification of Reviews using KNN: After the features are extracted then the final classification is done using KNN algorithm. This method is non-parametric, it takes K closest training set as input and class membership as output [21]. For this the whole data set is divided into training and testing dataset then the classification algorithm is applied on training dataset. For this the value of K (number of nearest neighbors) is decided then based on this the distance between the new data point and the K number of nearest

neighbors is calculated which can be Euclidian or Manhattan distance. After this calculation voting is done for example the two distance calculations say that it belongs to class positive, one say that it belongs to class negative then result will be that this new data point belongs to class positive .

Experimental Results: After classification step, experimental results are calculated in the form of precision, recall, f1-score, accuracy and execution time. If that accuracy is acceptable then this procedure will be applied for testing dataset and if not then the whole procedure is repeated again.

Classification for Prediction: When the classification is done then output will be like how many reviews are positive, negative and neutral and this can be used for analyzing that how many consumers are satisfied or not satisfied with that product.

Results: Output of these reviews can be used by vendors for their business when they will come to know the whole scenario about their consumers whether they are satisfied or not so that they improve their services, quality, and policies.

V. EXPERIMENTAL RESULTS

Dataset Description: Dataset contains the reviews of consumer of Amazon website collected from UCI repository. It contains 1963 reviews of consumers [20] regarding Amazon products. The whole dataset is divided in training and testing dataset out of which 60% is training and 40% is testing.

Classification Results: Results of both classifiers are measured and compared in terms of Accuracy, Execution Time, Precision, Recall, f1-score. Classification report contains measures such as precision, recall and f1-score. There are three classes for classification that are positive (-1), neutral (0) and negative (+1).

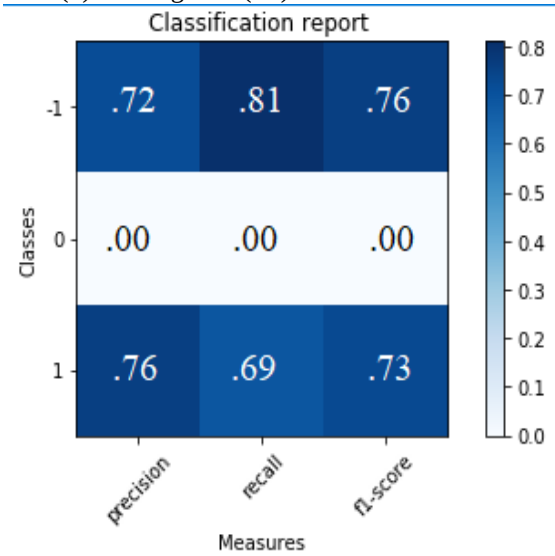

Fig. 3. CLASSIFICATION REPORT OF NAIVE BAYES

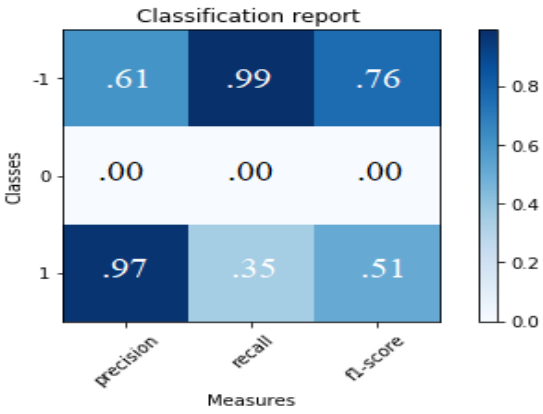

Fig.4.CLASSIFICATION REPORT OF HYBRID CLASSIFIER

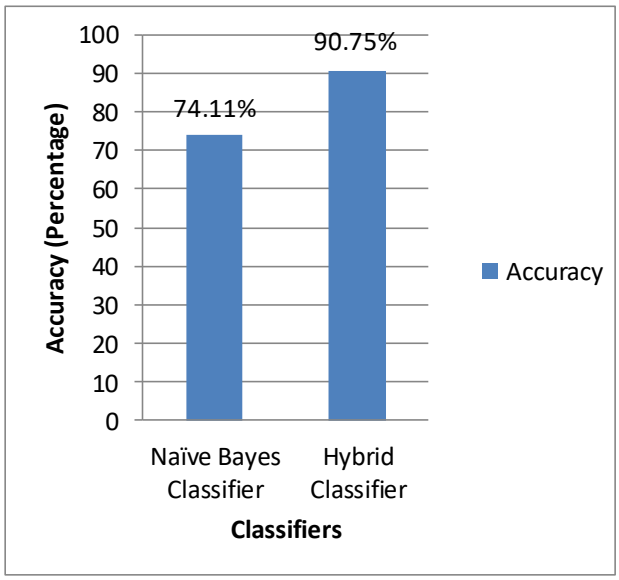

Fig. 5. ACCURACY COMPARISON

As shown in fig.5. The accuracy of the Naïve Bayes is compared with the Hybrid classifier. The Hybrid classifier is having high accuracy as compared to Naïve Bayes classifier. Naïve Bayes classifier is having 74.11% accuracy while on the other hand hybrid classifier is having 90.75% accuracy.

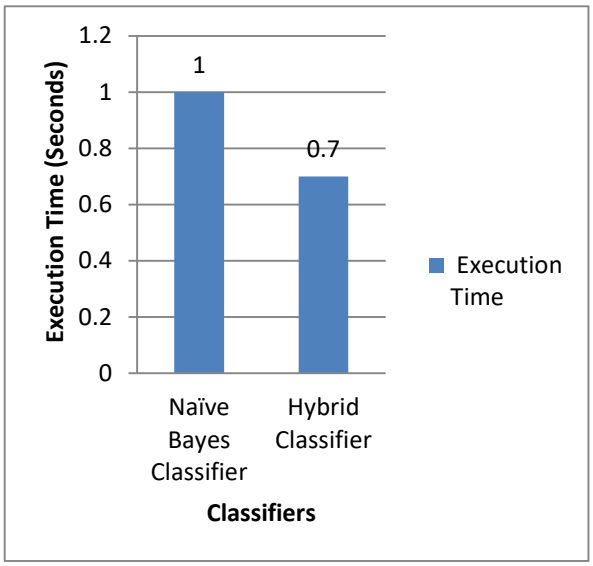

Fig. 6. EXECUTION TIME COMPARISON

In fig. 6. The comparative analysis of execution time of Naïve Bayes and hybrid classifier is shown. The Naïve Bayes classifier has taken execution time of 1 second and the hybrid classifier has taken .7 second for execution.

VI. CONCLUSION

The prediction analysis is an approach of machine learning based on training the models then calculating its accuracy, if that accuracy is acceptable then that model will be used for predicting future possibilities based on the past information. This research work is based on the prediction analysis in which customer behavior prediction is done using e-commerce consumer reviews of products using hybrid classification approach considering Decision Tree and KNN. In the previous research, naïve Bayes classifier was applied for the customer behavior analysis but the accuracy was low to improve the accuracy of naïve Bayes, the hybrid classification approach is designed and implemented in this work. It is observed that the accuracy of the hybrid classification is more as when compared to naïve Bayes classifier. For testing the proposed approach several performance metrics are used precision, recall, f1 score and execution time. As consumer behavior is an emerging field, this proposed approach can be used in many fields as accuracy is high and for the future work apart from having only consumer reviews other types of data can be used for predicting consumer behavior which will be very beneficial.

REFERENCES

1. W. Seo, et al., "Product opportunity identification based on internal capabilities using text mining and association rule mining," Technological Forecasting and Social Change, vol. 105, pp. 94- 104, 2016.
2. A. W. Joshi and S. Sharma, "Customer knowledge development: antecedents and impact on new product performance," Journal of Marketing, vol. 68, pp. 47-59, 2004.
3. K. Goffin and C. New, "Customer support and new product development-An exploratory study," International Journal of Operations & Production Management, vol. 21, pp. 275-301, 2001.
4. S. Tuarob and C. S. Tucker, "Quantifying product favorability and extracting notable product features using large scale social media data," Journal of Computing and Information Science in Engineering, vol. 15, p. 031003, 2015.
5. M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments," Expert Systems with Applications, vol. 40, pp. 4241-4251, 2013.
6. Gemma A. Calvert, Michael J. Brammer, "Predicting Consumer Behavior: Using Novel Mind-Reading Approaches", IEEE Pulse, 2012, Volume: 3, Issue: 3, Pages: 38 – 41.
7. A.C.M. Fong, Baoyao Zhou, Siu Hui, Jie Tang, Guan Hong, "Generation of Personalized Ontology Based on Consumer Emotion and Behavior Analysis", IEEE Transactions on Affective Computing, 2012, Volume: 3, Issue: 2, Page s: 152 – 164.
8. X. G. Luo, C. K. Kwong, J. F. Tang, F. Q. Sun, "QFD-Based Product Planning With Consumer Choice Analysis", IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2015, Volume: 45, Issue: 3, Pages: 454 – 461.
9. Lianju Ning, Haoyu Wang, Xin Feng, Junping Du, "The Browsing Pattern and Review Model of Online Consumers Based on Large Data Analysis", Chinese Journal of Electronics, 2015, Volume: 24, Issue: 1, Pages: 58 – 64.
10. Jungwoo Shin, Manseok Jo, Jongsu Lee, Daeho Lee, "Strategic Management of Cloud Computing Services: Focusing on Consumer Adoption Behavior", IEEE Transactions on Engineering Management, 2014, Volume: 61 , Issue: 3, Pages: 419 – 427.
11. Ru Jia, Ru Li, Meiju Yu, Shanshan Wang, "E-commerce Purchase Prediction Approach By User Behavior Data", 2017, International Conference on Computer, Information and Telecommunication Systems (CITS).
12. Yusheng Zhou and Shuiqing Yang, "Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews", IEEE Access, 2019, Volume: 7, Pages: 27769 – 27780.
13. Ting Bai, Wanye Xin Zhao, Yulan He, Jian-Yun Nie Ji-Rong Wen, "Characterizing and Predicting Early Reviewers for Effective Product Marketing on E-Commerce Websites", IEEE Transactions on Knowledge and Data Engineering, 2018, Volume: 30, Issue: 12, Pages: 2271 – 2284.
14. Namuk Ko, Byeongki Jeong, Sungchul Choi and Janghyeok Yoon, "Identifying Product Opportunities Using Social Media Mining: Application of Topic Modeling and Chance Discovery Theory", IEEE Access, 2018, Volume: 6, Pages: 1680 – 1693.
15. Yuanlin Chen, Yueting Chai, Yi Liu, and Yang Xu, "Analysis of Review Helpfulness Based on Consumer Perspective", Tsinghua Science and Technology, 2015, Volume: 20, Issue: 3, Page s: 293 – 305.
16. Hernandez Sergio, Pedro Alvarez, Javier Fabra and Joaquın Ezpeleta, "Analysis of users' behaviour in structured e-commerce websites", IEEE Access, 2017, Volume: 5, Pages: 11941 – 11958.
17. J. Tuladhar, A. Gupta, S. Shrestha, U. Bania and K. Bhargavi, "Predictive Analysis of E-Commerce Products", *Intelligent Computing and Information and Communication*, pp. 279-289, 2018.
18. N. Midha and V. Singh, "Classification of E-commerce Products Using RepTree and K-means Hybrid Approach", *Advances in Intelligent Systems and Computing*, pp. 265-273, 2017.
19. C. Troussas, A. Krouska and M. Virvou, "Trends on Sentiment Analysis over Social Networks: Pre-processing Ramifications, Stand-Alone Classifiers and Ensemble Averaging", *Machine Learning Paradigms*, pp. 161-186, 2018.
20. R. Ireland and A. Liu, "Application of data analytics for product design: Sentiment analysis of online product reviews", *CIRP Journal of Manufacturing Science and Technology*, vol. 23, pp. 128-144, 2018.
21. K. Srujan, S. Nikhil, H. Raghav Rao, K. Karthik, B. Harish and H. Keerthi Kumar, "Classification of Amazon Book Reviews Based on Sentiment Analysis", *Advances in Intelligent Systems and Computing*, pp. 401-411, 2018.
22. R. Jagdale, V. Shirsat and S. Deshmukh, "Sentiment Analysis on Product Reviews Using Machine Learning Techniques", *Cognitive Informatics and Soft Computing*, pp. 639-647, 2018.
23. R. McCarthy, M. McCarthy, W. Ceccucci and L. Halawi, "Introduction to Predictive Analytics", *Applying Predictive Analytics*, pp. 1-25, 2019.

AUTHORS PROFILE

Ms. Kareena is pursuing as a student of Master's of Engineering in Computer Science from Chandigarh University**.** She has done Bachelor's of Technology from T.I.T&S Bhiwani. Her area of interest is Data Mining and Machine Learning.

Er. Raj Kumar is working as an Assistant Professor in Computer Science Department of Chandigarh University. He has done M.Tech from CDLU Sirsa, MCA from KUK and pursuing PhD from IKGPTU Jalandhar. He has 21 papers published in international journals.