Impact of Feature Selection Technique on Email Classification

Aakanksha Sharaff, Naresh Kumar Nagwani, and Kunal Swami

***Abstract*—Being one of the most powerful and fastest way of communication, the popularity of email has led to untoward rise of email spam. Spam are unwanted and unsolicited messages and the subsequent rise of spam received by email users has become a serious security threat. Automatic filtering of spam emails, hence, is a promising and research worthy area whereupon extensive work has been reported about attempts to design machine learning based classifiers. Herein feature selection technique can be conveniently applied for developing efficient machine learning based classifiers. However, feature selection techniques provide a mechanism to identify suitable and relevant features (attributes) for any knowledge discovery task. The choice of selecting a suitable feature selection technique is always a key question of research. The present paper compares and discusses the effectiveness of two feature selection methods i.e. Chi-square and Info-gain on machine learning techniques namely Bayes algorithm, tree-based algorithm and support vector machine with a purpose to design a classifier for spam email filtering. The experiment is performed using 10-fold cross-validation and performance measures such as accuracy, precision, recall are used to compare the results.**

***Index Terms*—Classification algorithms, email spam Filtering, feature selection.**

I. Introduction

Email is a fastest, cheapest and efficient way of communication. Therefore spammers prefer to send spam through email. Spam is defined as Unsolicited Bulk Email (UBE). Spam is a big problem as it not only takes recipients time and wastes network bandwidth but also floods the mailboxes leaving them unmanaged. This increases the chances of missing an important email which can cause a serious problem to users. The delivery of legitimate emails is also affected by large amount of spam-traffic.

According to Cisco 2014 Annual Security Report, although the spam volume was on a downward trend worldwide in 2013, the proportion of maliciously intended spam remained constant. Spammers prey on people's desire for more information in the wake of a major event and trick them into a desired action, such as clicking an infected link. Thus, spam is not only an irritating factor but also a serious security threat. This brings in a great need for spam filters which can automatically filter spams and clean our mailboxes. The initial spam filters required a user to create rules. These rules were mainly designed by observing patterns in a typical spam email such as presence of specific words, combinations of words, phrases, etc. To bypass these rules, spammers employed content obfuscation techniques. For example, splitting or modifying words, such as 'lottery' written as '|0ttery'. Machine learning (ML) algorithms, which analyze the content of a message, have been successfully used to filter email spam. Supervised machine learning methods extract knowledge from training datasets supplied and use the obtained information to classify newly received email messages. ML algorithms automate the process of spam filtering and thus don't require users to be tech-savvy to create rules explicitly. Further, these algorithms are adaptive in nature and thus adapt to new and changing nature of spam.

Various ML classification algorithms such as Naïve Bayesian classifiers, decision tree (such as J48), Support Vector Machine (SVM), etc. have been successfully used to solve the problem of spam filtering. Along with these classification algorithms, the use of feature selection techniques also varied. There is an ample choice for classification and feature selection algorithms for spam filtering task. The aim of this paper is to investigate the effectiveness of various feature selection methods on different classification algorithms. In this study, three types of classifiers, Naïve Bayes classifier, Support Vector Machines (SVM) and J48 have been tested on a well-known publicly available data set, Ling-Spam [1] in conjunction with two feature selection techniques i.e. Chi-Square and Information Gain. Chi-Square [2] is a statistical measure of divergence from the expected frequency assuming the feature occurrence is actually independent of the class value. Information Gain [2] represents number of information bits for prediction of class of an attribute in a document. It is also denoted as Entropy (H) which is a measure of impurity that helps to decide the interestingness of a feature and ensures how much the feature is relevant for classification of training data.

The rest of the paper is organized as follows. Section II summarizes related work, Section III describes the overall approach, Section IV describes the result analysis and performance measures used and Section V concludes the paper.

II. Related Work

As found in literature, the first known machine learning approach to filter spam emails was proposed by using Naïve Bayes as classifier [3]. The authors incorporated several hand-crafted phrasal features, non-textual domain specific features and used Mutual Information feature selection technique. A series of experiments were performed using

Manuscript received September 14, 2014; revised February 4, 2015.

Aakanksha Sharaff and Naresh Kumar Nagwani are with the Department of Computer Science & Engg., National Institute of Technology, Raipur, 492010, India (e-mail: asharaff.cs@nitrr.ac.in, nknagwani.cs@nitrr.ac.in).

Kunal Swami is with Samsung Research India, India.

DOI: 10.7763/IJKE.2015.V1.10

Naïve Bayes as a classifier by Androutsopoulos *et al*. The authors carried out a comparison of Naïve Bayes and a memory based classifier, studied the effect of attribute size, training-corpus size and effect of lemmatization and stopping using cost sensitive evaluation measures [4]. Hovold used word-position-based variant of Naïve Bayes [5]. Five different versions of Naïve Bayes algorithm were compared on six datasets in which Flexible Bayes and Multinomial Naïve Bayes gave collectively best results [6].

Islam *et al*. proposed a model of spam filter using both linear and non-linear SVM for tackling text and image based spam using appropriate kernel function [7]. The algorithms based on decision trees have also been used for the purpose of spam filtering. In [8], the author compared three decision tree classifiers namely, Naïve Bayes Tree (NBT), J48 and Logistic Model Tree (LMT) and it is shown that LMT performed best in terms of accuracy and false positive rate. J48 turned out to be the best classifier in terms of training time. A Genetic Algorithm-Support Vector Machine (GA-SVM) feature selection technique is developed in [9] which showed significant improvements over SVM in terms of classification accuracy and computation time.

Many comparative studies comparing the performances of several classification algorithms can also be found in the literature. Youn & McLeod compared four classifiers namely Neural Network, SVM, Naïve Bayes and J48 and studied the effect of data size and feature size on classification results [10]. J48 turned out to be the best classifier in terms of accuracy, precision and recall [10]. Lai compared Naïve Bayes, kNN, SVM and tf-idf with SVM, called as integrated approach. Comparison was performed using different parts of email namely, header, subject, body and all (combining all three) and the best result was obtained with SVM when features from all three parts were used [11]. Only one performance measure, accuracy was considered by the authors to compare the classifiers. Awad & Elseuofi compared Naïve Bayes, kNN, Artificial Neural Networks (ANN), SVM, Artificial Immune System and Rough Sets classification algorithms. Naïve Bayes turned out to be the best classifier. The variation of accuracy, precision and recall with different number of features was not studied by the authors [12]. Another work by Kumar *et al*. compared various classification algorithms and effect of various feature selection techniques namely, Fisher, Relief, Runs Filtering and Stepwise Discriminant Analysis. Fisher and Runs filtering feature selection techniques gave better performance than other feature selection techniques and Random Tree classification algorithm in conjunction with Fisher filtering gave best result in terms of accuracy [13]. Trivedi & Dey compared Genetic Search and Greedy Stepwise Search feature selection techniques in conjuction with Naïve Bayes, SVM and Genetic Algorithm classifiers. Greedy Stepwise Search gave good results and SVM turned out to be best classifier in terms of accuracy and false-positive rate [14].

III. The Approach

This paper compares the machine learning classification techniques namely Bayes algorithm, tree-based algorithm J48 and support vector machine with feature selection techniques to design a classifier for spam filtering. The input documents are taken from the well known dataset Ling spam which are pre-processed. Ling-Spam contains a total of 2893 emails out of which 481 (16.7%) are spam and 2412 (83.3%) are legitimate. The experiment is carried out with the help of WEKA [15], an open source data mining tool. WEKA has a bulk collection of machine learning algorithms made for data mining tasks. The Email dataset goes through pre-processing to generate Term Document Matrix. A feature vector space is created with the matrix. The elimination of irrelevant attributes from the training dataset reduces the dimension and only the informative words are taken into account for classification. It can be done by stopping (removing pronouns, prepositions etc.) and stemming (grouping of words from the same root word). The feature selection techniques Chi-square and Information Gain are applied over the feature vector in conjunction with classification algorithms. The whole process is described with the help of Fig. 1.

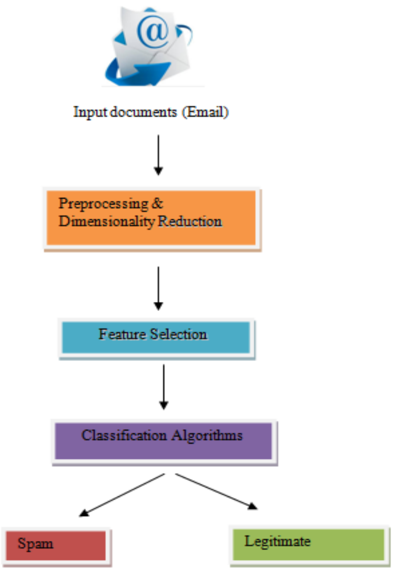
Input documents (Email)
Preprocessing & Dimensionality Reduction
Feature Selection
Classification Algorithms
Spam
Legitimate

Fig. 1. The overall approach.

A. Feature Selection Techniques

1) Chi-Square (χ^2)

Chi-Square is a statistical measure of divergence from the expected frequency assuming the feature occurrence is actually independent of the class value. The evaluation of this technique is performed by calculating the chi-square statistic with respect to the class of the attribute. For an initial hypothesis H_0, Chi-squared statistic is calculated as:

$$\chi 2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} \quad (1)$$

Here, O_{ij} is the observed frequency and E_{ij} is the expected frequency. The greater the value of χ^2, the greater is the evidence against the hypothesis H_0 [2].

2) Information Gain

Information Gain represents number of information bits for prediction of class of an attribute in a document. It is also denoted as Entropy (H) which is a measure of impurity that helps to decide the interestingness of a feature and ensures how much the feature is relevant for classification of training data [2]. The entropy of the dataset is calculated using the given formula:

$$H\left(D\right)=-\sum_{i=1}^{n}\left(\frac{D_i}{D}\right)log_2\left(\frac{D_i}{D}\right) \qquad (2)$$

The entropy of any subset is calculated as:

$$H\left(D|X\right)=-\sum_{j=1}^{v}\left(\frac{D_j}{D}\right)?H\ D\ X-X_j \qquad (3)$$

Here, $H(D|X-X_j)$ is the entropy calculated relative to the subset of instances that have a value of X_j for attribute X and *v* is the number of distinct values for attribute *X*. The term $\left(\frac{D_j}{D}\right)$ represents the weight of the *j*th partition.

The information gain of an attribute is measured by the reduction in entropies:

$$IG\left(X\right)=H\left(D\right)-H(D\,|\,X) \qquad (4)$$

IV. Results and Performance Analysis

The experiment was performed on the Ling-Spam dataset by applying the classification algorithms. The effect of feature selection techniques on these classification algorithms was also studied and the performance of the classification algorithms and their combinations with feature selection techniques is compared using the performance metrics.

For measuring the performance of the combinations certain popular performance metrics are used: Accuracy, Precision and Recall. These measures can be easily calculated with the confusion matrix. True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) are the 4 components of a confusion matrix as shown in Table I.

TABLE I: Confusion Matrix

		Predicted Class	
		Spam	Legitimate
Actual Class	Spam	TP	FN
	Legitimate	FP	TN

Taking spam as the positive class and legitimate as the negative class, the components can be defined as follows:

True Positive (TP): It is measured by number of spam emails correctly classified as spam.

True Negative (TN): It is measured by number of legitimate emails correctly classified as legitimate.

False Positive (FP): It is measured by number of legitimate emails incorrectly classified as spam.

False Negative (FN): It is measured by number of spam emails incorrectly classified as legitimate.

1) **Accuracy**: It is the ratio of emails correctly classified to the total number of emails.

$$Accuracy=\frac{\text{No.of emails correctly classified}}{\text{Total no.of emails}}$$

2) **Spam Precision**: It denotes the number of spams correctly classified to the total messages classified as spam.

$$\text{SP}=\frac{\text{No.of emails correctly classified as spam}}{\text{Total no.of emails classified as spam}}$$

3) **Spam Recall:** It is the percentage of all spams that are correctly classified as spam.

$$\text{SR}=\frac{\text{No.of emails correctly classified as spam}}{\text{Total no.of spam emails}}$$

A 10-fold cross validation technique is applied for calculation of metrics in which the dataset is broken into 10 parts. Out of 10, training is performed on 9 parts and 1 part of the set is used as a test. This method is repeated 10 times and finally the mean of the parameters are calculated.

Table II shows the formula for different performance measures used in this study. Accuracy gives the overall correctness of the model but not useful for analysing positive and negative correctness individually. Precision is used to determine the relevant positive classes out of total positives. Recall is the percentage of positives that are correctly detected as positive class.

TABLE II: Performance Measures

Performance Measure	**Formula**
Accuracy	$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$P=\frac{TP}{TP+FP}$
Recall	$R=\frac{TP}{TP+FN}$

A few observations can be made from the experiment. As shown in Fig. 2, the best classifier in terms of accuracy is SVM. The best accuracy of SVM was 99.34% and was obtained when all features were taken into account. With feature selection techniques, the accuracy of the SVM classifier is slightly decreased. This shows that SVM performs better without any feature selection technique. The next classifier with higher accuracy is Na ïve Bayes. Na ïve Bayes like SVM remains immune to feature selection techniques. The best accuracy obtained with Na ïve Bayes is ~97.8% and remains unaffected even with use of any feature selection technique. The next classifier in terms of accuracy is J48. J48 reports slight improvement in classification accuracy when feature selection techniques are employed. J48 reports an accuracy of ~96.7% and when feature selection techniques are employed, a slight improvement by ~0.5% is obtained.

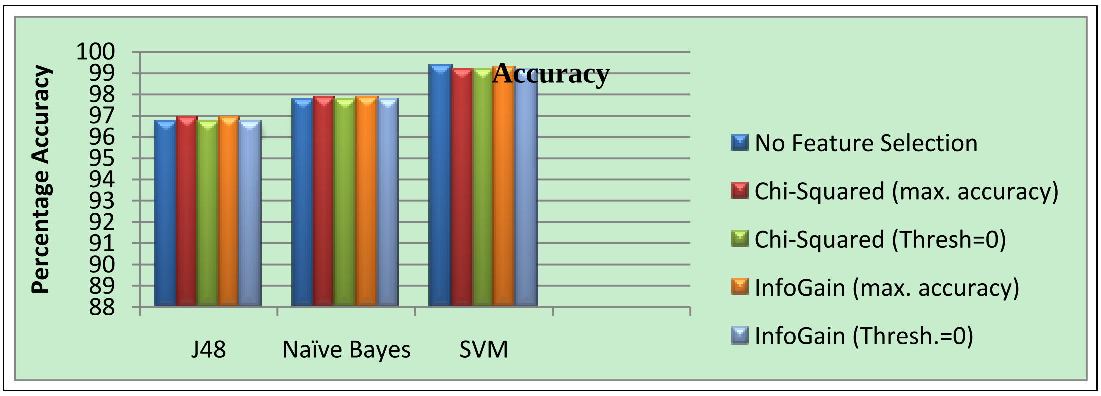

Fig. 2. Spam accuracy.

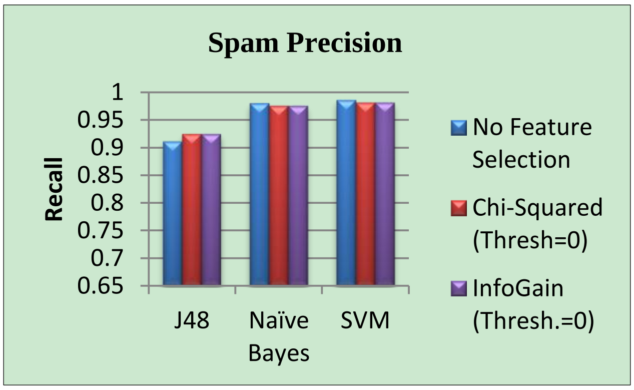

Fig. 3. Spam precision.

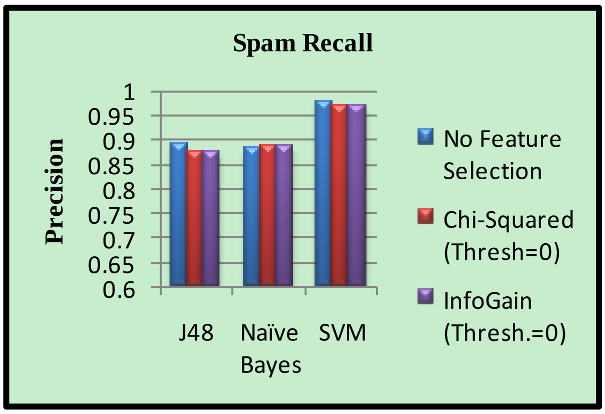

Fig. 4. Spam recall.

Effect of Feature Selection Technique on Accuracy

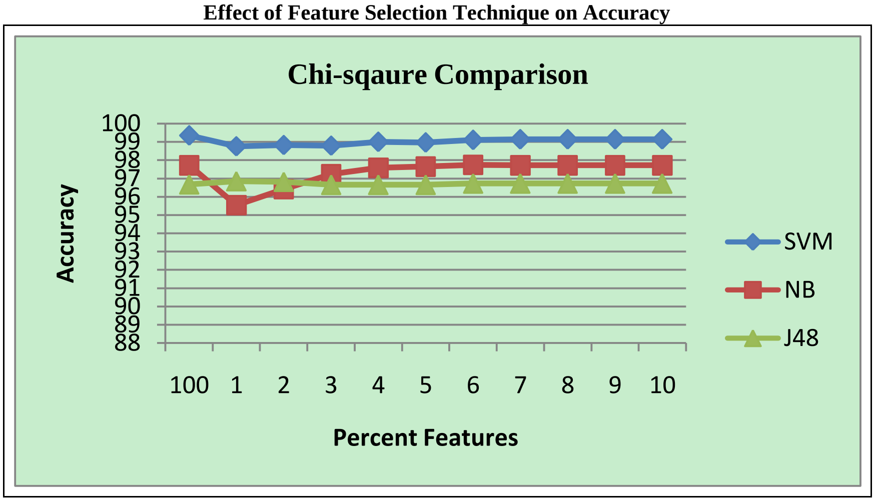

Fig. 5. Effect of Chi-square feature selection over accuracy.

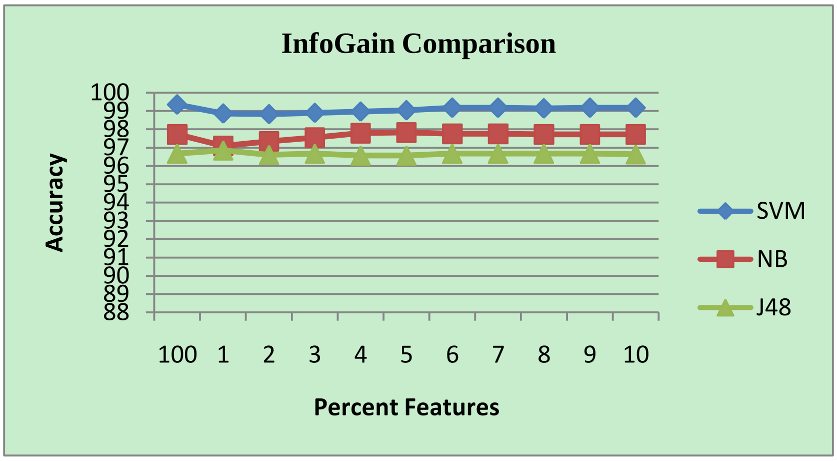

Fig. 6. Effect of Info-Gain feature selection over accuracy.

Following observation can be made from Fig. 3. The best classifier in terms of Spam Precision is SVM with consistent 98% precision in all cases. The next best classifier is Na ïve Bayes with consistent 97.5% precision in all cases. Next classifier is J48 whose precision is improved by more than 1% when feature selection is employed, precision goes over 92% with feature selection.

Following observation can be made from Fig. 4. Again SVM turns out to be the best classifier in terms of Spam Recall with ~97% which again remains nearly unaffected with feature selection techniques. Next best performers are Na ïve Bayes and J48 with ~88% spam recall.

The following results can be inferred from Fig. 5 and Fig. 6 that SVM is the best classifier among all the three when no feature selection technique is applied. Na ïve Bayes classifier performs better when Info- Gain feature selection technique is used and gives an accuracy of 97.71%.

V. Conclusion and Future Work

Among all classification technique SVM is the best performer and gives overall best results without employing any feature selection techniques. Na ïve Bayes and J48 turn out to be most consistent and good classification algorithm and like SVM, Na ïve Bayes has no significant effect of feature selection techniques. J48 shows slight improvement with feature selection. Among feature selection techniques, Info-Gain performs better compared to Chi-square feature selection technique.

In future we can compare more number of classification algorithms. An ensemble of SVM or Na ïve Bayes can be created and evaluated.

Acknowledgment

Authors want to thank National Institute of Technology, Raipur, India for its kind co-operation and support.

References

[1] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering," in *Proc. Workshop on Machine Learning in the New Information Age*, pp. 9-17, 2000.

[2] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav Journal of Operations Research*, vol. 21, pp. 119-135, 2011.

[3] M. Sahami, S. Dumais, D. Heckerman, and E.Horvitz, "A bayesian approach to filtering junk e-mail," In *Proc. the AAAI Workshop on Learning for Text Categorization*, 1998.

[4] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and C. D. Spyropoulos, "An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages," in *Proc. SIGIR 2000*, pp. 160-167, 2000.

[5] J. Hovold, "Naive bayes spam filtering using word-position-based attributes," in *Proc. the Second Conference on Email and Anti-Spam, CEAS*, 2005.

[6] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes - Which naive bayes?" in *Proc. Third Conference on Email and Anti-Spam (CEAS)*, pp. 125-134, 2006.

[7] M. R. Islam, M. U. Chowdhury, and W. Zhou, "An innovative spam filtering model based on support vector machine," in *Proc. the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)*, pp. 348-353, 2005.

[8] S. Chakraborty and B. Mondal, "Spam mail filtering technique using different decision tree classifiers through data mining approach - A comparative performance analysis," *International Journal of Computer Applications*, pp. 26-31, 2012.

[9] F. Temitayo, O. Stephen, and A. Abimbola, "Hybrid GA-SVM for efficient feature selection in e-mail classification," *Computer Engineering and Intelligent Systems*, vol. 3, no. 3, pp. 17-28, 2012.

[10] S. Youn and D. McLeod, "A comparative study for email classification," presented at International Joint Conferences on Computer, Information, System Sciences, and Engineering (CISSE06), 2006.

[11] C. Lai, "An empirical study of three machine learning methods for spam filtering," *Knowledge-Based Systems*, vol. 20, no. 3, pp. 249-254, 2007.

[12] W. A. Awad and S. M. Elseuofi, "Machine learning methods for e-mail classification," *International Journal of Computer Applications*, vol. 16, no. 1, pp. 39-45, 2011.

[13] R. K. Kumar, G. Poonkuzhali, and P. Sudhakar, "Comparative study on email spam classifier using data mining techniques," presented at the International MultiConference of Engineers and Computer Scientists (IMECS '12), Hong Kong, 2012.

[14] S. K. Trivedi and S. Dey, "Effect of feature selection methods on machine learning classifiers for detecting email spams," in *Proc.* the *ACM-SIGAPP Research in Adaptive and Convergent Systems (ACM RACS 2013)*, pp. 35-40, 2013.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

Aakanksha Sharaff has completed her graduation in computer science and engineering in 2010 from Government Engineering College, Bilaspur (C.G.). She has completed her post graduation master of technology in 2012 in computer science & engineering (specialization- software engineering) from National Institute of Technology, Rourkela and now she is pursuing the Ph.D. degree in computer science & engineering from National Institute of Technology Raipur, India. Her areas of interests are software engineering, data mining, text mining and information retrieval. She is currently working as an assistant professor at NIT Raipur India.

Naresh Kumar Nagwani has completed his graduation in computer science and engineering in 2001 from G. G. Central University, Bilaspur. He completed his post-graduation master of technology in information technology from ABV-Indian Institute of Information Technology, Gwalior in 2005 and completed the Ph.D. in computer science and engineering in 2013 from National Institute of Technology Raipur, India. His areas of interests are data mining, text mining, mining software repositories and information retrieval. His employment experience includes software developer and team lead at Persistent Systems Limited and presently he is an assistant professor at NIT Raipur. He has published more than 20 research papers in various journals and conferences.

Kunal Swami received his bachelor's degree in computer science & engineering in 2014 from National Institute of Technology, Raipur. Now, he is working as a software engineer at Samsung Research India.