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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

Enhancing Machine Based Spam Detection Using Twitter

Shinde Asha Ashokrao[1], Shital Y. Gaikwad[2]

M.E Second Year Student, Dept. of Computer Science and Engineering, Matoshri Pratishthan Group of Institutions Vishnupuri, Nanded.(M.S),India[1]

Asst.Professor.(B.E.M.Tech.), Dept. of Computer Science and Engineering, Matoshri Pratishthan Group of Institutions Vishnupuri, Nanded.(M.S),India[2]

ABSTRACT: As World Wide Web is expanding day by day, Tweet appear to be a reliable form of communication and the fastest way to send information from one place to another. Nowadays most of the transactions, be it general or business are taking place using Tweet as their mode of communication. Twitter is completely effective solution for communication as it helps in real time communication thereby saving time and expense. Apart from their advantages, Tweet are also affected by attacks that include spam tweets. Spam tweet are generally used to send bulk tweet to a sender. Spam floods the Internet with many similar copies of messages that is distributed in depth; these messages are sent strongly to recipients who would not otherwise choose to receive it. We will analyze several methods of mining data for spam data in order to figure out the best classifier for sorting Tweet. As part of this paper, we explain the classification of Tweet to identify spam and not spam. For this purpose, we use the Naive Bayesian Classifier and created an Tweets classification system to classify spam and not spam.

KEYWORDS: Tweets spam, Classification, Feature Extraction, Naive Bayesian Classifier

I. INTRODUCTION

Tweets is an efficient mode of online communication because it actually saves a expenses and helps in reducing time taken for communication thus making it a favorite communication medium in personal communication as good as in professional or business communication. Tweet always provide easy transfer information including textual and other files that can be sent around the world. There are more instances where the Tweet that we sent are affected by a series of attacks; these can be active or passive. Sometimes we receive Tweet from sources that are not known to us and few Tweet are also made up of irrelevant content that does not even matters to the user. These kinds of illicit and not required tweet are popularly known as Spam Tweet. Spam is the known practice of constantly sending unwanted or huge data to specific or random set of Tweets accounts. Spam Tweets is a subset of online spam involving completely identical or partially matching messages sent to various recipients via Tweets. Spam Tweet also include some malware in the form of scripts or other files that can be executable and that can harm user's system. Tweet and their so called spam lists are mostly created by thorough scanning Usenet ads, by stealing the Internet tweetsing list. Spam tweets has become a growing problem in recent years. It has been estimated that about 70% of all Tweet are spam. Just like web expansion, the problem of spam Tweet is also expanding at large. According to [14], it was found average span of 10 days every year is wasted in processing of spam.

Spam tweets has become a growing problem in recent years. It has been estimated that about 70% of all Tweet are spam. Just like web expansion, the problem of spam Tweet is also expanding at large. According to [13], it was found average span of 10 days every year is wasted in processing of spam.
Spam is also an expensive problem that can costs lots of expenses in a year to the service providers for the loss of bandwidth. Spam is also a major problem that also attacks the very existence of Tweet. Therefore, it is very essential to distinguish spam Tweet, many methods have been suggested to identify and classify Tweets messages as spam or non spam or legitimate tweets and it has been found that the algorithm success rate of machine learning is

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

extremely high. Several algorithms that are processed for the classification of unwanted Tweet that are widely used and analyzed amongst them are vector machine, Naive Bayes, decision tree, neural network classifiers are well-known classifiers. In this article, we experiment with algorithms: Naive Bayes, Bayes Net, Vector Machine Support (SVM), Tree Function (FT), J48, Random Forest and Random Tree.
In order to process and analyze the result that the classifiers works quite well with certain selected attributes, we need to apply a feature selection algorithm on to the same data set (the algorithm that we have used here is the First Match algorithm) and we can apply the same to classifiers and entities selected. With the help of this study, we can surely find that accuracy of each classifier has tremendously improved as and when we select the features by our algorithm Best-First again compared to all of the classifiers we have experienced in this dataset reduced Naive Bayes shows better results in the Context of precision.

II. BACKGROUND STUDY

This section provides an overview of what Data Mining is, a different algorithm for data mining, explains the selection of features and most of the terms we used in this document.

A. Data Mining

Data Mining is nothing but the discovery of the large database. It is a technique that tries to find new models in huge and heavy datasets. It is a mixture of different fields like artificial intelligence, machine learning, statistics and database systems. The main purpose of the data mining method is to extract information from a dataset and transform it into a form that is understandable for future use. In fact the original data mining task is the auto or semi-automatic analysis of large amounts of data to extract interesting patterns previously indefinite. While large-scale information technology has evolved from separate transactions and analytical systems, data mining provides a link between the two methods. Data mining software analyzes the relationships in the stored data based on end user questions. Basically Data mining uses basic steps as listed in five points:

1. Obtain, transform and process, and load data on the data warehouse system.
2. Save, store and process data in multidimensional database system.
3. Provide easier data access to the business analyst and also to the technical professionals.
4. Analyze and process data by using the existing tools and application softwares.
5. Convert data in appropriate format that can be used by the user to work with pictorial representations such as graphs or tables.

Sometimes we assume data mining is synonymous with another known term, Knowledge Discovery from databases (KDD), because data mining is a necessary step in the discovery process. Knowledge Detection is a combination of all steps illustrated in Figure 1.

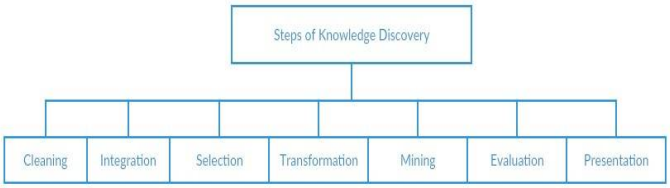

Figure1. Knowledge Discovery steps

B. Algorithm Used in this Study

Naive Bayes: A Naive Bayes classifier is a simple probabilistic classifier with strong assumptions of independence. Simply put, a naive bayes classifier assumes that the presence / absence of a particular property of a class is not related to the presence / absence of any other feature, considering the class variable as a function of the Class Probability Model, Trained in a supervised learning environment. An advantage of the naive bayes classification is that it requires only a small amount of training data to estimate the parameters required for classification. The

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

Bayesian classification assumes that the data belongs to a particular class. We then calculate the probability that the assumption is true. Bayesian classmates are basically statistical classifiers, that is, they can predict probabilities of class membership, such as the probability that a given test belongs to a particular class.

The naive technique of Bayes is based on a Bayesian approach; it is therefore a simple, clear and fast classifier [10]. Before reaching the main term of the Baye theorem, we will first analyze certain terms used in the theorem. P (A) is the probability that event A occurs. P (A / B) is the probability that event A occurs because event B has already occurred or we can define it as the conditional probability of A as a function of the condition that B has already occurred. The Bayes theorem is defined in Equation 1.

P (A/B) = P (B/A) P (A) P(B) (1)

If we consider OBJX as an object to be classified with the likelihood of belonging to one of classes CLS1, CLS2, CLS3, etc and calculating,

Prob (CLSi / OBJX). Once these probabilities have been calculated for all classes, we simply assign OBJX to the class with the highest probability.

Prob (CLSi / OBJX) = [Prob (OBJX / CLSi) Prob (CLSi)] / Prob (OBJX) (2)

Where Prob (CLSi / OBJX) is the probability that the object OBJX belongs to a class CLSi, Prob (OBJX / CLSi) is the probability of obtaining attribute values OBJX if we know that it belongs to class CLSi Prob (CLSi) is the probability of an object belonging to a class CLSi without any other information, and Prob (OBJX) is the probability of obtaining OBJX attribute values regardless of the class to which the object belongs.

C. Classification and Prediction

Classification is the separation of objects into classes. If classes are created without looking at the data, the classification is known as a priori classification. If classes are created by looking at the data, the classification method is known as the later classification. When classifying, it is assumed that the classes have been considered a priori and the classification then consists in forming the system so that when a new object is introduced into the formed system it can affect the object to one of the existing classes. This approach is popularly known as the supervised learning process where in meaningful inputs are given to the system to learn for iterative process. Data classification can be divided into a two-step process as depicted the figure given below (refer Figure 2). In the first step, the model is constructed describing a predetermined set of data classes. The model is constructed by analyzing the database tuples described by the attributes. It is assumed that each tuple belongs to one of the existing classes, as determined by the class label attribute. The data tuples analyzed to construct the model collectively form the training set.

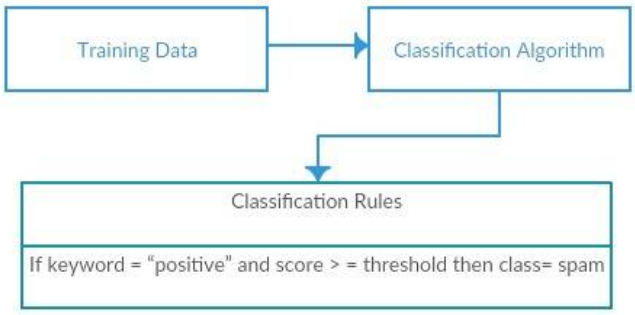

Figure 2: Learning and Training of Classifier

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

Table 1: Different Tweets Ids

Tweets ID	Type	Keywords	Class Label
@spammerGuy	Human	Credit, Dollars	Spam
@MyMaiden	Bot	Weight loss, Herbal	Spam
@Hasbro	Cyborg	Lottery, Polling Results	Spam

The second stage, as shown in figure 3, the model is used for classification. The predictive accuracy of the model is evaluated first. The accuracy of a model in a given test data set is the percentage of samples of test sets correctly classified by the model. For each test sample, the known class tag is compared to the class prediction of the model learned for that sample.

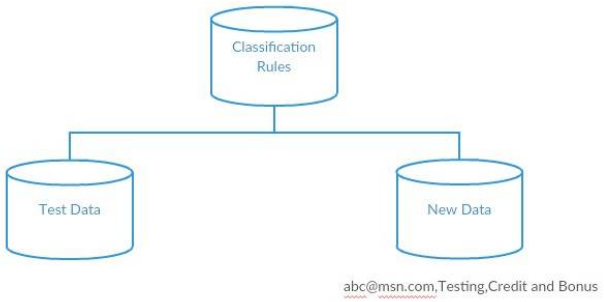

Figure 3: Classification Model

The prediction can be considered as the construction and use of a model to evaluate the class of an unmarked sample or to evaluate the ranges of values of an attribute that a given sample is likely to have. In this context, classification and regression are the two main types of prediction problems, where classification is used to predict discrete or nominal values, while regression is used to predict continuous or ordered values.

III. RELATED WORK

Spam tweet are one of the major problem areas in the Internet world that can bring economic losses to organizations and also bring harm to individual users as well. Tweets spam is also known as junk tweets that is sent to a group of recipients who have not requested it. Spam is a serious problem that threatens the existence of Tweets services. Since it is not a cost, it is cheap to send bulk Tweets to a group of users. It takes a lot of time to delete or sort these unwanted Tweet and also introduces a risk of deleting normal tweet by mistake. In the study [8] Rambow et al Application of machine learning techniques for Tweets summary, In this study, RIPPER classifier is used for determining the sentences to be included in a summary. The learning model uses features such as linguistic functionality, Tweets functions, and thread structure. This approach requires positive examples in large numbers and it is also found that abstracts do not occur for a variable length depending on the interest of the users.

There are so many techniques available to detect these unwanted Tweet. These approaches come mainly from the field of artificial intelligence, data mining or automatic learning. Automatic learning techniques are more varied and

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

widely used for spam classification. Decision tree classify spam using previous data [12]. But it is costly to calculate and recalculate that spammers change the technique. In the study [6] Bayesian networks found as the very popular technique for detecting junk tweets. But with this approach, it is quiet difficult to scale on many features to come out with judgment.

We have all been aware that spam tweet generate many problems in today's world. Therefore, several approaches are developed to stop spam. The primary purpose of anti-spam filtering is automatically excluding unwanted Tweet from user tweetsbox. These are desirable causes for problems such as populating tweetsboxes, engulfing important personal tweets, wasting a lot of network bandwidth also causes problems of congestion, time and loss of energy for users while sorting these Tweet undesirable effects. In the study [5], two methods are described for classification. First is done with certain rules that are defined manually, such as the expert system based on the rule. This classification technique applies when classes are static and their components are easily separated according to the characteristics.
Secondly it is done with the help of existing machine learning techniques. According to the study [1] Spam Tweets clusters are created with the help of function criteria. The criterion function is defined as the maximization of similarity between messages in clusters and this similarity is calculated using the nearest k-neighbor algorithm.

Symbiotic Data Mining is a distributed data mining approach that unifies the content-based filtering with collaborative filtering described in [7]. The main objective is to reuse local filters to improve personalized filtering in the context of privacy. In the study [15], tweets sorters based on the network feed approach of the neural network feedback and Bayes classifiers are evaluated. From this study, it can be seen that the forward and backward propagation neural network classifier provides very high accuracy over other existing classifiers. In the work [9], the Bayesian approach is applied to the problem of classification and grouping using a model based on assumptions such as population, subject, latent variable and sampling scheme.

According to [2], content filtering was one of the first types of anti-spam filter. These types of filters use coded rules that have an associated score and are updated periodically. A good example of this type of filter is Spam Assassin [3], which works by scanning the Tweets text document against each rule and adding points for all matching rules. According to the study [4], if the total Tweets score exceeds a certain threshold threshold, then this message falls into the spam category. To generate these scores, a single perception is used where perceptual inputs indicate whether a rule has been matched and the weight of the corresponding entry indicates the score for each rule.
In this article [11] spam is detected using the Naive Bayes Claissfier and hybrid techniques. In each classification, a custom threshold adjustment value can be calculated by comparing the actual output value with the expected output value. To detect unwanted and unwanted Tweets messages and prevent these unwanted messages from being passed to a user inbox is called a spam filter. The spam filter is a program like other types of filtering programs that look for certain criteria on which to base the trials. The entry of Tweets filtering software is Tweets. The unchanged message for delivery to the user's tweetsbox is the tweets filter output. Some of the tweets filters are able to edit messages during processing. Tweets filters have different degrees of configurability. From time to time, they focus on the choices taking into account the coordination of a coherent expression. Different times, the essential words in the body of the message are used, or perhaps the Tweets address of the sender of the message.

IV. PROPOSED WORK

Description: In this paper we are describing the method that is used to perform spam Tweets classification. The first step is to select the data set file and apply the feature extraction technique for the extracted feature. For which we are using the word-count algorithm. The next step is to form the set of data that is extracted using the characteristic extraction technique. For the formation of the data we can calculate the probability of spam and not spam words in the document. The next step is to test the data with the help of Naïve Bayesian Classifier for which it calculates the probability of spam and non-spam tweets and make a prediction whose value is greater. If spam words are larger than words that are not spam in an Tweets, the tweets is unwanted Tweet.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

In the next step we are calculating the words that are misclassified by the classifier and we calculate the accuracy of the classifier and also calculate the classifier error rate by calculating the fraction of the word that is misclassified and the total number of words in the document.

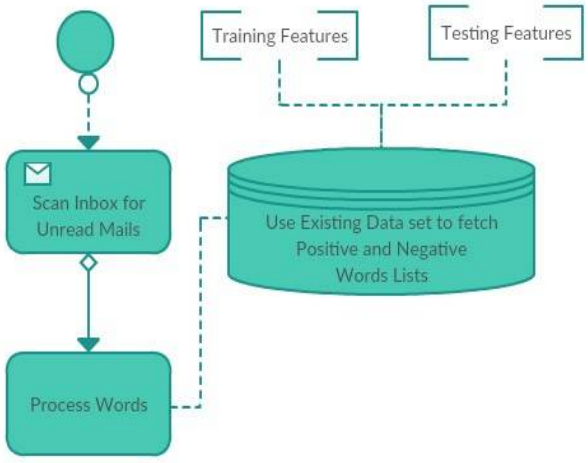

Figure 4: Word processing and classification for training using extract dataset

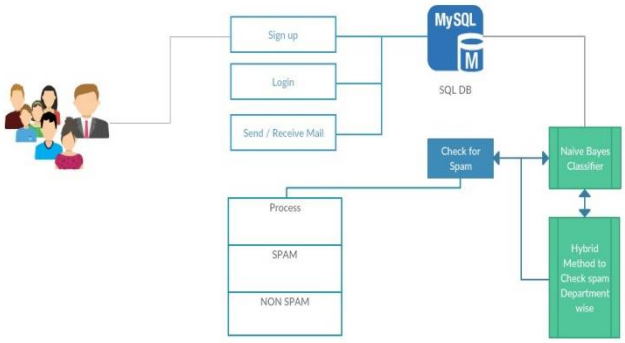

Figure 5: Proposed System

Proposed System:

Step 1: Select the Tweets

Step 2: Extract features with help of tokenization and word count algorithm.

Step 3: Training the dataset with the help of Naive Bayesian Classifier.

Step4: Find the probability of spam and non-spam tweet. Prob_spam = (sum(train_matrix(spam_indices,)) + 1) ./ (spam_wc + numtokens)

Prob_nonspam = (sum(train_matrix(nonspam_indices,)) + 1) ./ (nonspam_wc + numtokens)

Step 5: Testing the dataset

log_a = test_matrix*(log(prob_tokens_spam))' + log(prob_spam)

log_b = test_matrix*(log(prob_tokens_nonspam))'+ log(1 - prob_spam)

if output = log_a > log_b then document are spam else the document are non-spam

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

Step 6: Classify the spam and non-spam tweet.
Step 7: compute the error of the text data and calculate the word which is wrongly classified
Numdocs_wrong = sum(xor(output, text_lables))
Step 8: display the error rate of text data and calculate the fraction of wrongly classified word
Fraction_wrong = numdocs_wrong/numtest_docs

V. RESULTS & DISCUSSION

As part of this project, we explain the classification of Tweet to identify spam and not spam. For this purpose, we use the Naive Bayesian Classifier and created an Tweets classification system to classify spam and not spam. To do this, we created a custom dataset to run this experiment. In our data set, we took a total of 960 Tweet in which 700 train data and 260 test data. Of the 700 data streams, 350 are spam and 350 are non-spam. Similarly, the test data set 260 contains 130 spam Tweet and 130 non-spam Tweet.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$FMeasure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Here we are present reading different for the four sets of data formed that are tested by the classifier, ie naive Bayesian Classifier and Support Vector Machine. Hence the different readings and calculation of the result:

Table 2: Readings of different Classifiers

Train Dataset	Support Vector Machine	Naive Bayes Classifier
Dataset 100	68	78.66
Dataset 200	66.23	79.12
Dataset 700	65.11	81.34

This reading contains the text data classified by the classifier and provides the word that is misclassified and the classifier error rate. Therefore, we can show the overall result provided by the classifier. Hence Naive Bayesian classifier classifies most of the words accurately. When the number of data sets is increased, the Bayesian naive classifier produces a better result compared to the vector machine.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

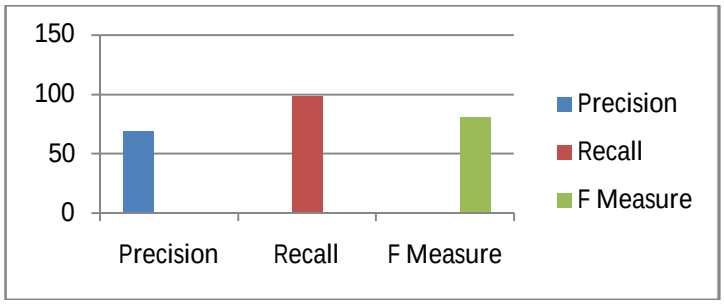

Figure 6: Graphical Representation of Performance Metrics

Therefore, we can show the graph to display the accuracy and error rate of both the Naive Bayesian Classifier and Vector Machine

VI. CONCLUSION

In the classification of spam, the main source of concern is the classification of Tweets and unwanted threats. So today, most researchers are working in this area to find the best classifier to detect spam. Therefore, you need a filter with great precision to filter out spam tweet or spam tweet. In this article, we have focused on finding the best classifier for spam classification using data mining techniques. Therefore, we apply different classification algorithms in the given input data set and verify the results. In this study, we analyzed that classifiers work well when we incorporate the feature selection approach into the classification process, that is, accuracy is dramatically improved when 0063lassifiers are applied to the classifier, Dataset rather than the data set.

We use the Bayesian Naïve Classifier here and extract the word using the word count algorithm. After computing, we find that the naive Bayesian classifier has the most accurate vector machine. The error rate is very low when we use the Bayesian Naïve Classifier. It can be said that Naïve Bayesian Classifier produces a better result than Support Vector Machine.

REFERENCES

[1] Clark J. et al (2010), "A Neural Network Based Approach to Automated Tweets Classification.

[2] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R.Strachan,"Hybriddecision tree and naïve Bayes classifiers for multi-class classification tasks,"Expert Syst.Appl.,vol.41,no.4 PART 2, pp. 1937– 1946,2014.

[3] D. T. Larose, Data Mining Methods and Models. John Wiley & Sons, 2006.

[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 Tweets Messages," Proc.23rd Annu.Int. ACM SIGIR Conf. Res.Dev. Inf. Retr., pp. 24–28, 2000.

[5] I. Idris, A.Selamat, and S. Omatu, "Hybrid Tweets spam detection model with negative selection algorithm and differential evolution," Eng. Appl. Artif. Intell., vol. 28, pp. 97–110, Feb.2014.

[6] I. Idris and A. Selamat, "Improved Tweets spam detection model with negative selection algorithm and particle swarm optimization," Appl. Soft Comput., vol.22, pp. 11–27, 2014.

[7] I. Idris, A. Selamat, N. Thanh Nguyen, S. Omatu, O. Krejcar, K. Kuca, and M. Penhaker, "A combined negative selection algorithm-particle swarm optimization for an Tweets spam detection system," Eng. Appl. Artif. Intell., vol. 39, pp. 33–44, 2015.

[8] Jia Z. et al (2012) "Research on Web Spam Detection Base on Support Vector Machine"IEEE 2012 pp. 517-520.

[9] J.R.Quinlan, "J. Ross Quinlan_C4.5_ Programs for Machine Learning.pdf," Morgan Kaufmann, vol. 5, no. 3. p. 302, 1993.

[10] Liu B. et al (2013) "Scalable Sentiment Classification for Big Data Analysis Using Na¨ıve Bayes Classifier" IEEE 2013 pp.99-104.

[11] M. Lichman, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]." Irvine,CA: University of California, School of Information and Computer Science.

[12] Panigrahi P. (2012), "A Comparative Study of Supervised Machine Learning Techniques for Spam Tweets Filtering", Fourth International Conference on Computational Intelligence and Communication Networks, IEEE 2012, pp. 506-512.

[13] Sharma A. and Anchal (2014), "SMS Spam Detection Using Neural Network Classifier",ISSN: 2277 128X Volume 4, Issue 6,

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 8, August 2017

June 2014, pp. 240-244.
[14] Sharma K. and Jatana N. (2014)"Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach" IEEE 2014 pp.939-942.
[15] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," Knowledge- Based Syst., vol. 64, pp. 22–31, 2014.

BIOGRAPHY

Shinde Asha Ashokrao is a second year student of master of engineering, Dept. of Computer Science and Engineering, Matoshri Pratishthan Group of Institutions Vishnupuri, Nanded.(M.S),India.

Ms Shital Y Gaikwad is a Assistant Professor, department of computer science and engineering, Matoshri Pratishthan Group of Institutions Vishnupuri, Nanded.(M.S),India. She has received Master of technology degree from SRTMU, Nanded.