Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

Enhancing The Spam Detection Technique Using Naïve Bayes Classifier Algorithm

Harpreet Kaur
Sri Guru Granth Sahib World University

Abstract— Due to the rapid growth of digital data made available in recent years, knowledge discovery and data mining have attracted a great deal of attention with an imminent need for turning such data into useful information and knowledge. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

Social networking is an emerging area for market purpose also. A large number of traffic is engaged in social media and gather around for particular discussions or reviews on a topic or a simple gadget. However spamming of comments are still under main problem which can affect the website's reputation as well as makes an ordinary user vulnerable to malware or spyware in their system. Numerous number of string search algorithms which are responsible for filtering out the spam comment are been studied and are needed to be improved. Various string search algorithms are K Means, KNN, AdaBoost Classifier, Decision stump, Decision tree, Genetic Algorithm, Naïve bayes, apriori algorithms etc. But the previous results were noticeable that the decision tree and naïve bayes algorithms performed little better as compared to another algorithms. Finally, best classifier for facebook spam comment is identified based on the Accuracy, Precision, F-Measure, Execution Time and Recall**.**

Keywords: Spam Detection, Naïve Bayes Algorithm and Decision Tree Algorithm.

I. INTRODUCTION

Data mining (knowledge discovery from data) is the process of extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data and transforms it into an understandable structure for further use. Alternative names of data mining are Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc. First international conference on Data Mining was held in 1995. Data Mining can be categorized in two subsequent ways that are classification and clustering. Data mining is the key process of knowledge discovery from database.

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data and may be used in further analysis.

In this paper focuses on a different classification techniques that are most commonly used in spam detection. The study is carried out on two algorithms (Naïve Bayes and Decision tree) to show the accuracy, precision, recall, f-measure and execution time of these two classification algorithms that used to spam detection. Finally, best classifier for facebook spam comment is identified based on the Accuracy, Precision, F-Measure, Execution Time and Recall**.**

II. RELATED WORKS

Sachin S. Patil et al have presented survey for various data mining techniques such as classification and clustering techniques. For detection of intrusions and web based attacks of and every activities such as ID3 is an classification algorithm used for constructing a decision tree, it is a predictive model to predict a unknown classifier, it is a greedy top down approach, there are many applications are proposed based on ID3. ID3 algorithm has some drawbacks as it produced decision trees with large no of anomalies or useless data due to which the decision tree built with many branches, and the information measure gain measure tends to prefer attributes with many values. Also once the decision tree is built it may contain useless rules. To overcome all these drawbacks of ID3 algorithm, it is the need of time to make improvement in ID3 algorithm, with the increasing use of network and database technologies various intrusion, anomalies attacks are possible now a days. With the use of internet web servers are popular targets of the attacks with HTTP request at web server. Hence further research is needed to use the improved ID3 algorithm for detection of web based attacks such as SQL attacks, anomalies, intrusions **[22].**

Ishtiaq Ahmed et al have proposed a hybrid system of SMS classification to detect spam or ham, using Naïve Bayes classifier and Apriori algorithm. Though this technique is fully logic based, its performance will rely on statistical character of the database. Naïve Bayes is considered as one of the most effectual and significant learning algorithms for machine learning and data mining and also has been treated as a core technique in information retrieval. However, by applying user-specified minimum support and minimum confidence, they gain significant improvement on effective accuracy 98.7% from the traditional Naïve Bayes approach 97.4 **[14].**

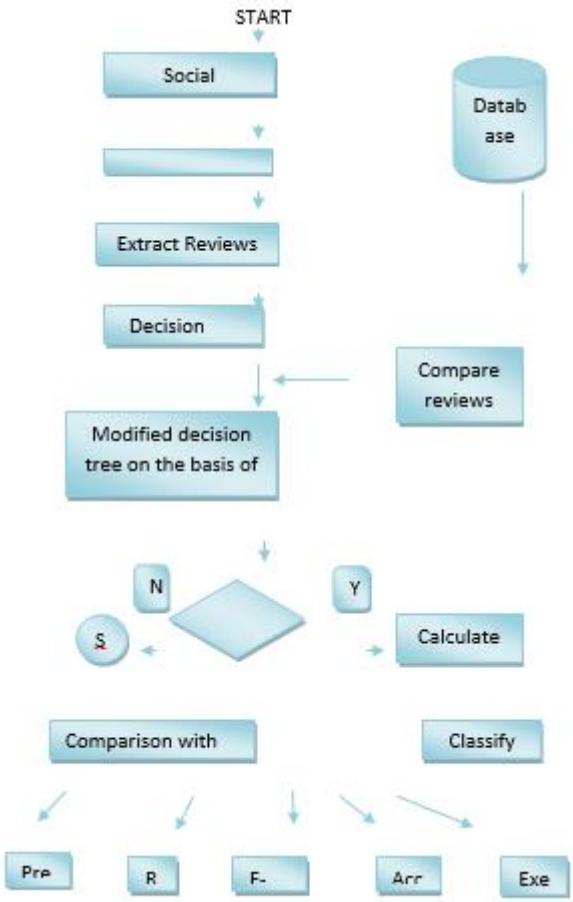
Daniela XHEMALI et al have described web classification has been attempted through many different technologies. In this study they concentrate on the comparison of Neural Networks (NN), Naïve Bayes (NB) and Decision Tree (DT) classifiers for the automatic analysis and classification of attribute data from training course web pages. They introduce an enhanced NB classifier and run the same data sample through the DT and NN classifiers to determine the success rate of our classifier in the training courses domain. They shows that enhanced NB classifier not only outperforms the traditional NB classifier, but also performs similarly as good, if not better, than some more popular, rival techniques. Also shows that, overall NB classifier is the best choice for the training courses domain, achieving an impressive F-Measure value of over 97%, despite it being trained with fewer samples than any of the classification systems have encountered **[13].**

Tina R. Patil et al have presented classification is an important data mining technique with broad applications to classify the various kinds of data used in nearly every field of our life. Classification is used to classify the item according to the features of the item with respect to the predefined set of classes. This paper put a light on performance evaluation based on the correct and incorrect instances of data classification using Naïve Bayes and J48 classification algorithm. Naive Bayes algorithm is based on probability and j48 algorithm is based on decision tree. The paper sets out to make comparative evaluation of classifiers Naïve bayes and J48 in the context of bank dataset to maximize true positive rate and minimize false positive rate of defaulters rather than achieving only higher classification

accuracy using WEKA tool. The experiments results shown in this paper are about classification accuracy, sensitivity and specificity. The results in the paper on this dataset also show that the efficiency and accuracy of j48 is better than that of Naïve bayes.

This proves that the, J48 is a simple classifier technique to make a decision tree. Efficient result has been taken from bank dataset using weka tool in the experiment. Naive Bayes classifier also showing good results. The experiments results shown in the study are about classification accuracy and cost analysis. J48 gives more classification accuracy for class mortgage in bank dataset having two values Yes and No. Though here in this example, cost analysis valued same for both the classifier, with gender attribute, we can prove that J48 is cost efficient than the Naive Bayes classifier **[24].**

III. FRAMEWORK OF PROPOSED SYSTEM



IV. STUDY OF CLASSIFICATION ALGORITHM

Data mining (knowledge discovery from data) is the process of extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data and transforms it into an understandable structure for further use. Data Mining can be categorized in two subsequent ways that are classification and clustering.

4.1 Algorithm Classification based on two different data mining techniques:

A) Association Rule: Association rule mining, one of the most important and well researched techniques of data mining. Its aims to finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transactional databases, relational databases, and other information repositories. A typical and widely-used example of association rule mining is Market Basket Analysis.

Association rule provide information of this type in the form of "If-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rule are probabilistic in nature.

B) Classification: Data Mining techniques such as classification, association and clustering are generally used to extract the hidden, previously unseen knowledge from luminous of databases. Of the various data analysis techniques, classification is a supervised machine learning techniques which makes predictions about the future class instances by mapping instances of testing data to the predefined class labels which is learnt from the supplied instances of classes with class labels. These are several models in classification such as probabilistic model and evolutionary algorithmic model etc. Data classification is a two-step process: **Learning step (or training phase):** where a classification model is constructed and **Classification step (or test phase):** where the model is used to predict class labels for given data.

Classification consists of assigning a class label to a set of unclassified cases:

- **Supervised learning:** In which the class label of each training tuple is provided, this step is also known as supervised learning (i.e. the learning of the classifier is "supervised" in that it is told to which class each training tuple belongs).
- **Unsupervised learning:** In which the class label of each training tuple is not known and the number or set of classes to be learned may not be known in advance.

4.2 Decision Tree

Decision tree is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. A decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. Assume that all of the features have finite discrete domains, and there is a single target feature called the classification. Each element of the domain of the classification is called a class.

A decision tree or a classification tree is a tree in which each internal node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes. A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data.

In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Types of Decision Tree

Decision trees used in data mining are of two main types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node.

Advantages of Decision Tree

- **Simple to understand and interpret.** People are able to understand decision tree models after a brief explanation.
- **Requires little data preparation:** Other techniques often require data normalization, dummy variables need to be created and blank values to be removed.
- **Able to handle both numerical and categorical data**: Other techniques are usually specialized in analyzing datasets that have only one type of variable.

There are many specific decision-tree algorithms. Some of them are ID3 (Iterative Dichotomiser 3), C4.5 (successor of ID3), CART (Classification And Regression Tree), CHAID (CHI-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees and MARS: extends decision trees to handle numerical data better.

4.3 Naïve Bayes Classifiers

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a Naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficiency of naive Bayes classifiers.

Still, a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests. An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Advantages:

- Fast to train (single scan).
- Fast to classify.
- Not sensitive to irrelevant features.
- Handles real and discrete data.
- Handles streaming data well.

Disadvantages:

- Assumes independence of features.

4.3.1 Uses of Naive Bayes Classification

A) Naive Bayes text classification

The Bayesian classification is used as a probabilistic learning method (Naive Bayes text classification). Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents.

B) Spam filtering

Spam filtering is the best known use of Naive Bayesian text classification. It makes use of a naive Bayes classifier to identify spam e-mail. Bayesian spam filtering has become a popular mechanism to distinguish illegitimate spam email from legitimate email .Many modern mail clients implement Bayesian spam filtering. Users can also install separate email filtering programs. Server-side email filters, such as DSPAM, Spam Assassin, Spam Bayes, Bogofilter and ASSP, make use of Bayesian spam filtering techniques, and the functionality is sometimes embedded within mail server software itself.

C) Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering

Recommender Systems apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. It is proposed a unique switching hybrid recommendation approach by combining a Naive Bayes classification approach with the collaborative filtering. Experimental results on two different data sets, show that the proposed algorithm is scalable and provide better performance–in terms of accuracy and coverage–than other algorithms while at the same time eliminates some recorded problems with the recommender systems.

D) Online applications

This online application has been set up as a simple example of supervised machine learning and affective computing.

V. EXPERIMENTAL RESULT AND PERFORMANCE EVOLUTION

The setup is run for gaining insight into the performance against the following parameters. Parameters that are used named as precision, recall, f-measure, accuracy, and execution time.

5.1 Precision: Precision is the ratio of positive words and total words in database.

$$P = \frac{\text{Positive words}}{\text{Total Words in Database}}$$

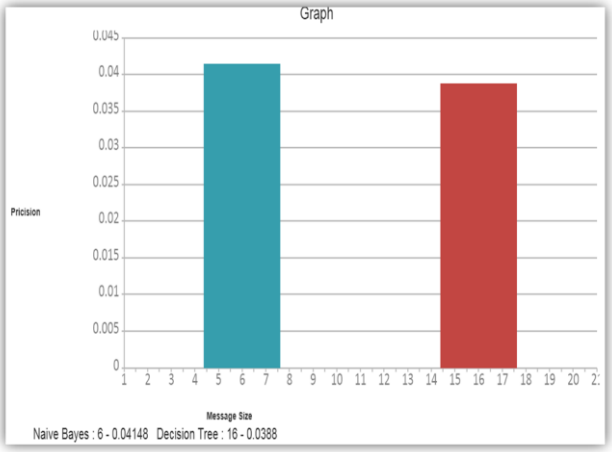


Fig 5.1 Comparison Graph With Respect To Precision

5.2 Recall: Recall is the ratio of positive word and positive words plus negative words.

$$R = \frac{\text{Positive words}}{\text{Positive words + Negative words}}$$

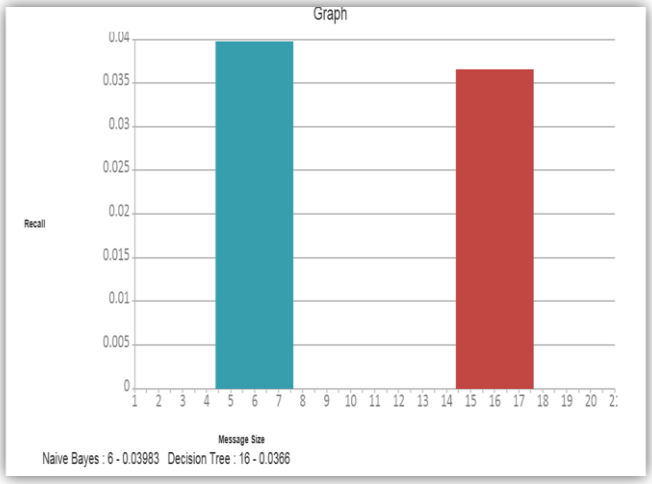


Fig. 5.2 Comparison Graph With Respect To Recall

5.3 F-Measure: F-measure is the harmonic mean of recall and precision.

$$\text{F-measure} = \frac{\text{Recall*Precision}}{\text{(Recall + Precision)}}$$

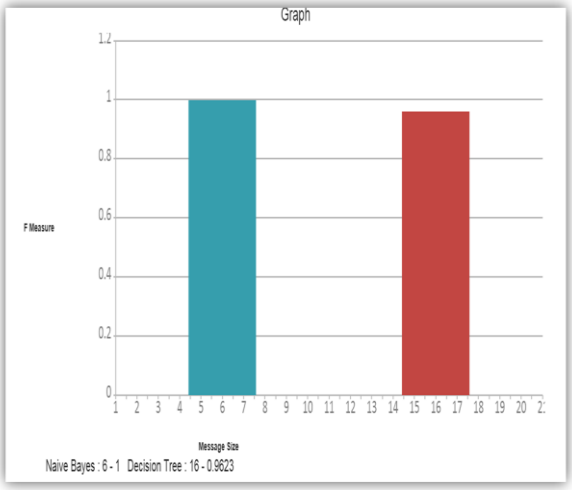


Fig. 5.3 Comparison Graph With Respect To F-Measure

5.4 Accuracy: Accuracy is defined as degree of conformity and correctness of something when compared to true or absolute value. The more common definition associates accuracy with systematic errors and precision with random errors. In other words accuracy can be defined as the combination of both trueness and precision. Accuracy is measured using the average of precision and recall.

$$\text{Accuracy} = \frac{\text{Precision} * (1 + \text{Recall})}{(\text{Recall} + \text{Precision})}$$

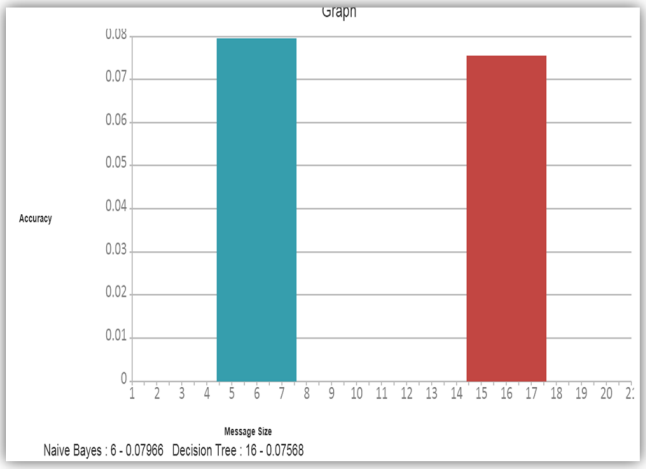


Fig. 5.4 Comparison Graph With Respect To Accuracy

5.5 Execution Time: Execution time is the time during which a program is running or executing. Following graph shows that execution time of decision tree is more as compared to naïve bayes.

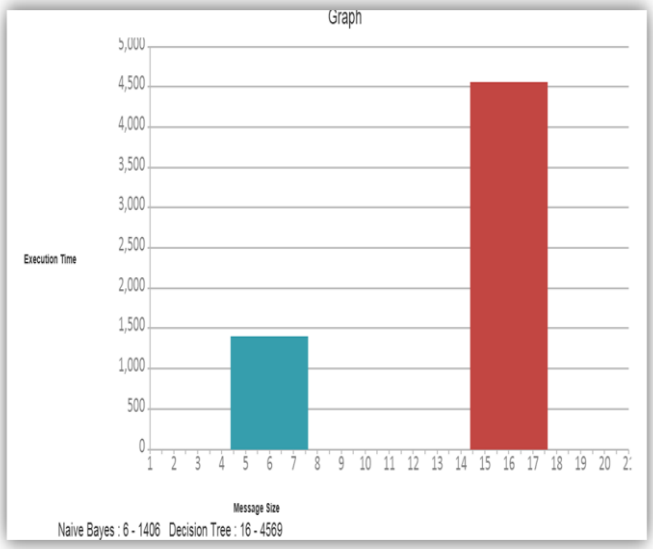


Fig. 5.5 Comparison Graph With Respect To Execution Time

Comparison Table

Table 5.6 Comparison Results of decision tree and naïve bayes classifier

Parameters	Decision tree values	Naïve bayes values
Precision	0.039124675	0.041484717
Recall	0.90851814	1.0
Execution Time	3628	2142
Accuracy	0.07805252	0.07966457
F-Measure	0.037162356	0.039832287
Message Count	6	6

VI. RESULT

The above graphs and tables show that naïve bayes classifier is best as compared to decision tree. Following are the reasons that show why naïve bayes classifier is best as compared to decision tree.
Naïve bayes is more accurate.
It provides more precise output.
F-measure and recall gives better output.
Naïve bayes can be easily implemented.
Naïve bayes does pick up the full string comments while Decision Tree does not pick up full string comments.

VII. CONCLUSION AND FUTURE WORK

Spam messages are nuisance and huge problem to most users since they clutter their mailboxes and waste their time to delete all the junk mails before reading the legitimate ones. They also cost user

money with dial up connections; waste network bandwidth and disk space. There are different existing algorithms to fight against spam. But the previous results were noticeable that the decision tree and naïve bayes algorithms performed little better as compared to another algorithms (like K Means, KNN, AdaBoost Classifier, Decision stump, Genetic Algorithm).Because both of these algorithms are simple to understand and interpret, requires little data preparation, able to handle both numerical and categorical data .

The research deals with the classification of the spam messages and detecting it. The implementation analyzes the two classification algorithm's naïve bayes and decision tree classifier working along with synonyms extraction. Naïve Bayes classifier is best as compared to decision tree because Naïve Bayes classifier is more accurate, provides more precise output, better recall and F-measure output and easy to implement. Naïve Bayes classifier does pick up the full string comments while decision tree does not pick up the full string comment.

VIII. FUTURE WORK

Classification is an important technique of data mining and has applications in various fields. In the present study few parameters like recall, precision, execution time, F-Measures and accuracy.
But there are still many parameters that can be taken into consideration for further research.

REFERENCES

[1] A. S. Galathiya, A. P. Ganatra et al;" Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning", International Journal of Computer Science and Information Technologies,, Volume 3, No.2, 2012.

[2] Anshul Goyal and Rajni Mehta et al;" Performance Comparison of Naïve Bayes and J48 Classification Algorithms", International Journal of Applied Engineering Research, Volume 7, No.11, March 2012.

[3] Ala' Eshmawi and Suku NairEddin et al;" Feature Reduction for Optimum SMS Spam Filtering Using Domain Knowledge", IJCSI International Journal of Computer Science Issues, Volume 10, No.2, March 2013.

[4] Ann Nosseir Khaled Nagati and Islam Taj-Eddin et al;" Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks", IJCSI International Journal of Computer Science Issues, Volume 10, No.2, March 2013.

[5] Amit Anand Soni, Abhishek Mathur;" Content based web spam detection using naive bayes with different feature representation technique ", Int. Journal of Engineering Research and Applications, Volume 3, No.5, Sept - oct 2013, pp.198-205.

[6] Ahmad Ashari, Donghai Guan et al;" Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool ", International Journal Of Computer Science And Applications, Volume 4, No.11, 2013.

[7] Bhawana S.Dakhare and Ujwala V.Gaikwad;"Spam Detection and Filtering using Different Methods ", International Journal of Computer Applications, No.1, 2012.

[8] Bhavesh Patankar and Dr. Vijay Chavda et al;" A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, No.22, December 2014

[9] C. Dellarocas;"Strategic manipulation of internet opinion forums: Implications for consumers and firms", Management Science, Volume 52, No.10, pp 1577-1593, 2006

[10] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee.;" How opinions are received by online communities: a case study on amazon.com helpfulness votes", In Proceedings Of the 18th WWW Conference, pp 141–150, 2009.

[11] C.L. Lai, K.Q. Xu, Raymond Y.K. Lau et al;" High-order Concept Associations Mining and Inferential Language Modeling for Online Review Spam Detection", IEEE International Conference on Data Mining Workshops, 2010.

[12] Congfu Xu(1), Baojun Su et al;" An Adaptive Fusion Algorithm for Spam Detection", IEEE International Conference on Data Mining Workshops, Volume 1, No.2, 2013.

[13] Daniela XHEMALI, Christopher J. HINDE et al;" Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages ", IJCSI International Journal of Computer Science Issues, Volume 4, No.2, 2013.

[14] Ishtiaq Ahmed, Donghai Guan et al;" SMS Classification Based on Naïve Bayes Classifier and Apriori Algorithm Frequent Itemset ", International Journal of Machine Learning and Computing, Volume 4, No.2, April 2014.
[15] K Butchi Raju and Chinta Someswara Rao;" Parallel String Matching Problems with Computing Models – An Analysis of the Most Recent Studies", International Journal of Computer Applications (0975 – 8887), Volume 13, No.11, August 2013.
[16] Manasi Kulkarni et al;" Web Spam Detection Using C5.0 Classification Algorithm ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, No.2, Feb 2013.
[17] N. Jindal and B. Liu;" Analyzing and detecting review spam", In Proceedings of the Seventh IEEE International Conference on Data Mining, pp 547–552, 2007.
[18] RasimM. Alguliev, Ramiz M. Aliguliyev, and Saadat A. Nazirova;" Classification of Textual E-Mail Spam Using Data Mining Techniques", Institute of Information Technology of Azerbaijan National Academy of Sciences,Volume10, No. 1155, 5 September 2011.
[19] R. Kishore Kumar, G. Poonkuzhaliet al;" Comparative Study on Email Spam Classifier using Data Mining Techniques", International Multiconference of Engineers and Computer Scientistis, Volume 1, No.14-12, 2012.
[20] Sarit Chakraborty, Bikromadittya Mondal;" Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis ", International Journal of Computer Science , Volume 47, No.16, June 2012.
[21] Sahil Puri, Dishant Gosain, Mehak Ahuja, Ishita Kathuria, Nishtha Jatana et;"Comparison and Analysis of Spam Detection Algorithms", IJCSI International Journal of Computer Science Issues, Volume 2, No.4, April 2013.
[22] Sachin S. Patil, Deepak Kapgate et al;" A Review on Detection of Web Based Attacks using Data Mining Techniques ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, No.12, December 2013.
[23] Sugandha Sharma et al;" E-Mail Spam Detection Using NLP ", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, No.6, june 2014.
[24] Tina R. Patil, Mrs. S. S. Sherekar et al;" Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification ", International Journal Of Computer Science And Applications, Volume 6, No.2, April 2014.
[25] Willian Hua et al;" Threshold and Associative Based Classification for Social Spam Profile Detection on Twitter ", IJCSI International Journal of Computer Science Issues, Volume 10, No.2, March 2013.

