PROJECT REPORT

ON

DETECTION OF SPAM MESSAGES ON SOCIAL NETWORKING SITES AND MAILS

SUBMITTED TO-

SHIKHA MEHTA

SHIKHA JAIN

SUBMITTED BY-

MAYANK GARG 13104768

VISHAL BISHT 13104751

B-11

INTRODUCTION

The spammers are getting smarter and the challenge of developing accurate spam filters is a big challenge. E-mail and social networking sites spam has steadily, even exponentially grown since the early 1990s to several billion messages in a single day. Spam has frustrated, confused, and annoyed e-mail and social networking sites users, by wasting time and valuable resources. E-mail provides a perfect way to send millions of advertisements at no cost for the sender, and this unfortunate fact is nowadays extensively exploited by several organizations. As a result, the e-mailboxes of millions of people get cluttered with all this so-called *unsolicited bulk e-mail* also known as "spam" or "junk mail". Large amounts of spam-traffic between servers cause delays in delivery of legitimate data, people with dial-up Internet access have to spend bandwidth downloading junk mail. Sorting out the unwanted messages takes time and introduces a risk of deleting normal mail by mistake. Finally, there is quite an amount of pornographic spam that should not be exposed to children. There are several ways to fight spam like legal measures, blocking spammer's IP-address and at last E-mail filtering using machine learning algorithms like Naïve Bayes Classifier, Decision Tree etc. which is the topic of further study.

LITERATURE SURVEY

Sources – IEEE, Microsoft Research, Books,

Title of paper	Implementing Spam Detection using Bayesian and Porter Stemmer Keyword Stripping Approaches
Authors	Biju Issac, Wendy J. Jap
Year of Publication	2009
Publishing details	TENCON 2009 - 2009 IEEE Region 10 Conference on 23-26 Jan, 2009
Summary	This paper talks about spam emails rise where one's email is bombarded with email that makes no sense at all. Firstly, Porter Stemmer Algorithm is discussed which deals with stemming the keywords to common root words like caring to care etc. according to 5 steps from step 1 to step 5. Bayesian based approach is used for further processing in which we train the data having both spam and non spam messages often called "ham" and the spam filter keeps record of the words which are used in both spam and ham emails. When the spam filter accepts test mail it extracts the keywords and gives probability score to every keyword based on the number of occurrences gathered during training stage. The probability score of each keyword will be summed up. The total score will be summed up and if it exceeds the threshold value we decide whether the email is spam or ham. Further for improvement keyword context matching is applied.
Web Link	http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5396056

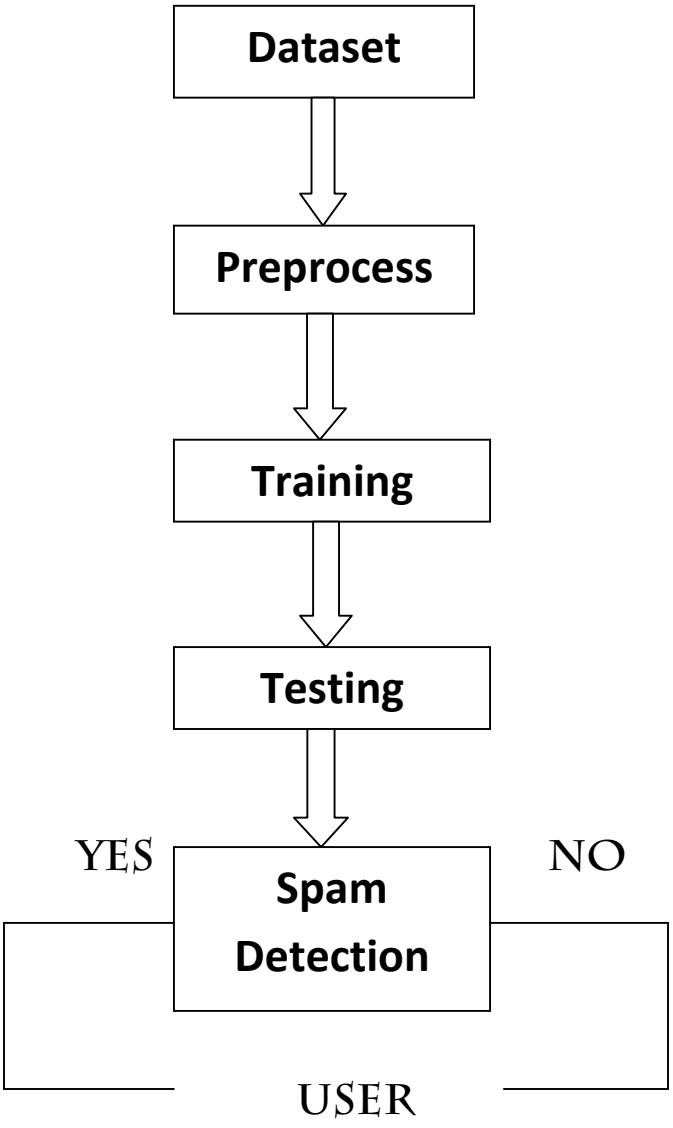
Title of paper	DON'T FOLLOW M E *Spam Detection in Twitter*
Authors	Alex Hai Wang
Year of Publication	2010
Publishing details	Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on 26-28 July, 2010
Summary	In this paper, a spam detection prototype system is proposed to identify suspicious users on Twitter. A directed social graph model is proposed to explore the "follower" and "friend" relationships among Twitter. Based on Twitter's spam policy, novel content-based features and graph-based features are also proposed to facilitate spam detection. A Web crawler is developed relying on API methods provided by Twitter. Around 25K users, 500K tweets, and 49M follower/friend relationships in total are collected from public available data on Twitter. Bayesian classification algorithm is applied to distinguish the suspicious behaviors from normal ones. Dataset is analyzed and performance of the detection system is evaluated. Classic evaluation metrics are used to compare the performance of various traditional classification methods. Experiment results show that the Bayesian classifier has the best overall performance in term of *F*-measure. The trained classifier is also applied to the entire data set. The result shows that the spam detection system can achieve 89% precision.
Web Link	http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5741690

Title of paper	Using a Data Mining Approach: Spam Detection on Facebook
Authors	M. Soiraya, S. Thanalerdmongkol, C. Chantrapornchai
Year of Publication	2012
Publishing details	International Journal of Computer Applications (0975 – 8887) Volume 58– No.13, November 2012
Summary	We develop an application to test the prototype of Facebook spam detection. In the development, we rely on the Facebook APIs for acquiring user data. The features for checking spams are the number of keywords, the average number of words, the text length, and the number of links. The Facebook APIs are web services of Facebook which can be called by Python client library. The important technology used is Facebook Graph API. Given the URL, it is searched in the blacklist database. If not found, the post text will be used next. Then Keyword blocking is done so as prohibited words are detected and access is rejected immediately. If not, preprocessing is needed after extracting the post text. Swath is used to chop words in Thai with the help of separators and dictionary. After that, the stopping words are eliminated to maintain only words that should be useful for detection. For testing the model, we divide the data into 2 sets. There are 150 posts for training which contains 75 normal posts and 75 spam posts. The testing set is 100 posts which contain 50 normal posts and 50 spam posts. The training data is fed into J48 model in Weka. Purposely it will create a decision tree. The testing data will test the accuracy of the decision tree made.
Web Link	http://research.ijcaonline.org/volume58/number13/pxc3883660.pdf

Title of paper	Detecting LinkedIn Spammers and its Spam Nets
Authors	Victor M. Prieto , Manuel Alvarez´ and Fidel Cacheda
Year of Publication	2013
Publishing details	(IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 4, No. 9
Summary	We propose a method for detecting Spammers and Spam nets in the LinkedIn social network. As there are no public or private LinkedIn datasets in the state of the art, we have manually built a dataset of real LinkedIn users, classifying them as Spammers or legitimate users. The data is then actually extracted from LinkedIn API of the profiles and data is then manually stored. The detection method we will proceed is with features such as: Number of words in profile, Number of contacts, Name size: in this case, the deficiency of Spam profiles appears in the name of the person. Fake profiles usually contain shorter names and surnames than in legitimate profiles, Location size: we have observed that Spam profiles usually contain a simple and smaller location than in the legitimate profiles, Name written in lowercase, Profile with photo, Plagiarism in profiles. To evaluate the classifier we used the Weka j48 technique. In each iteration a new model is built and assessed, using one of the sets as a “test set” and the rest as “training set”.
Web Link	https://thesai.org/Downloads/Volume4No9/Paper_30-Detecting_Linkedin_Spammersand_its_Spam_Nets.pdf

RESULTS OF LITERATURE SURVEY

SUMMARY



Emails today are a fast and inexpensive mode of sharing personal and business information in a convenient way. Often we find our inbox full of undesirable mails. So, it has become essential to have reliable tools to detect spam and ham mails. A naïve bayes filter which provides best results is applied on emails; if it is an unsolicited mail then it would be dropped to the junk folder else if ham (those mails which are sent by genuine users) then it would be dropped into the inbox. Even Twitter and other social networking sites are in full use today so there are also chances of spamming so Bayesian classifier is best suited to provide solution to this problem. The paper deals with the spam detection and improves its accuracy by using porter stemmer algorithm which reduces the word to its common root word so the processing is done only on different words and not on the same type of words. Second paper deals with the spam detection on twitter in which friends and followers directed social graph is build and data is collected using twitter API and web crawler and content based filtering is applied using Bayesian model which achieves an accuracy of 89% on the entire dataset and finally tells that Naïve Bayes Classifier gives best result. Third paper deals with Facebook spam detection. The features for checking spams are the number of keywords, the average number of words, the text length, and the number of links. Given the URL, it is searched in the blacklist database. If not found, the post text will be used next. Then Keyword blocking is done so as prohibited words are detected and access is rejected. If not, preprocessing is needed after extracting the post text. The stopping words are eliminated to maintain only words that should be useful for detection. For testing the model, we divide the data into 2 sets. The training data is fed into J48 model in Weka. Purposely it will create a decision tree. The testing data will test the accuracy of the decision tree made. The last paper enquires about detecting Spammers and Spam nets in the LinkedIn social network. Real LinkedIn users, classifying them as Spammers or legitimate users. The data is then actually extracted from LinkedIn API of the profiles. The detection method we will proceed is with features such as: Number of words in profile, Number of contacts, Name size: in this case, the deficiency of Spam profiles appears in

the name of the person. Fake profiles usually contain shorter names and surnames than in legitimate profiles, Location size: we have observed that Spam profiles usually contain a simple and smaller location than in the legitimate profiles, Name written in lowercase, Profile with photo, Plagiarism in profiles. To evaluate the classifier we used the Weka j48 technique.

OPEN PROBLEMS

Email spam and ham messages further categorization. Ham messages are further classified to primary and promotional whereas spam messages can be further classified to adult and non adult content. Using classifiers, comparative analysis will be done. If all the applied algorithms gives the same result then it would be considered as final else majority algorithms giving same results will be considered as final.

PROBLEM STATEMENT

Spam is one of the main problems of the WWW. Many studies exist about characterizing and detecting several types of Spam (mainly Web Spam, Email Spam, Forum/Blob Spam and Social Networking Spam). The purpose of this project is to detect spam messages in the mass generated messages. We detect spam messages and try to improve its runtime complexity and accuracy by comparing different classifier techniques of data mining such as Naïve Bayes, Decision Tree etc. Using classifiers, comparative analysis will be done. If all the applied algorithms gives the same result then it would be considered as final else majority algorithms giving same results will be considered as final.

OVERVIEW OF PROPOSED APPROACH

Firstly, we collect the dataset either from internet source or by using APIs of different social sites. Now, preprocessing is done over the dataset to speed up the algorithm. In preprocessing, stopword are removed as they only increase the dimension of the dataset and stemming is done to identify common root word which makes the results better for further NLP process. Data containing both spam and ham are divided into 70% and 30%. 70% data is trained such that the keywords are divided into positive or ham and negative or spam. Using different classifiers 30% data is tested and accuracy is checked. Then, test mails are classified as spam or ham and naïve bayes filter extracts the keywords and gives probability score to every keyword based on the number of occurrences gathered during training stage. The probability score of each keyword will be summed up. The total score will be summed up and if it exceeds the threshold value we decide whether the data is spam or ham. Decision Tree firstly decides according to which attribute the decision has to be made using entropy calculation and information gain. The attribute with maximum information gain is selected then if decision can’t be made further then the same process is applied. Using classifiers, comparative analysis will be done. If all the applied algorithms gives the same result then it would be considered as final else majority algorithms giving same results will be considered as final. Due to this, one will get to know about the spamming activity taking place in his/her account and about the spam messages which are of no use.

DESCRIPTION

Email and other social networking sites provide a perfect way to send millions of advertisement at no cost to the sender, and this fact is nowadays extensively exploited by several organizations. As a result, the e-mailboxes and social networking sites accounts of millions of people get flooded with unsolicited bulk messages also known as spam. To overcome these problems spam filters are being used to trap the spam messages beforehand. One such spam filter is made using Naïve Bayes Classifier a machine learning algorithm which does not require specifying any rules explicitly. Instead, training samples is needed. A specific algorithm is then used to "learn" the classification rules from this data. Data with known spam and ham messages is divided into some percentage and then one part of data is trained to create a bag of words of spam and ham keywords. Then, the other part of the data is tested and accuracy is measured and if our classifier works fine then it can further be applied to test messages which will produce results whether the given message is spam or ham. Naïve Bayes Classifier uses probabilistic distribution to classify the messages using prior probability and likelihood. Decision Tree firstly decides according to which attribute the decision has to be made using entropy calculation and information gain. The attribute with maximum information gain is selected then if decision can't be made further then the same process is applied. Using classifiers, comparative analysis will be done. If all the applied algorithms gives the same result then it would be considered as final else majority algorithms giving same results will be considered as final.

FUNCTIONAL REQUIREMENTS

Textual Dataset must be supplied.

Preprocessing, training and testing.

User entered data.

NON FUNCTIONAL REQUIREMENTS

Usability

The system is designed with completely automated process hence there is no or less user intervention.

Reliability

The system is more reliable because of the qualities inherited from the chosen platform python.

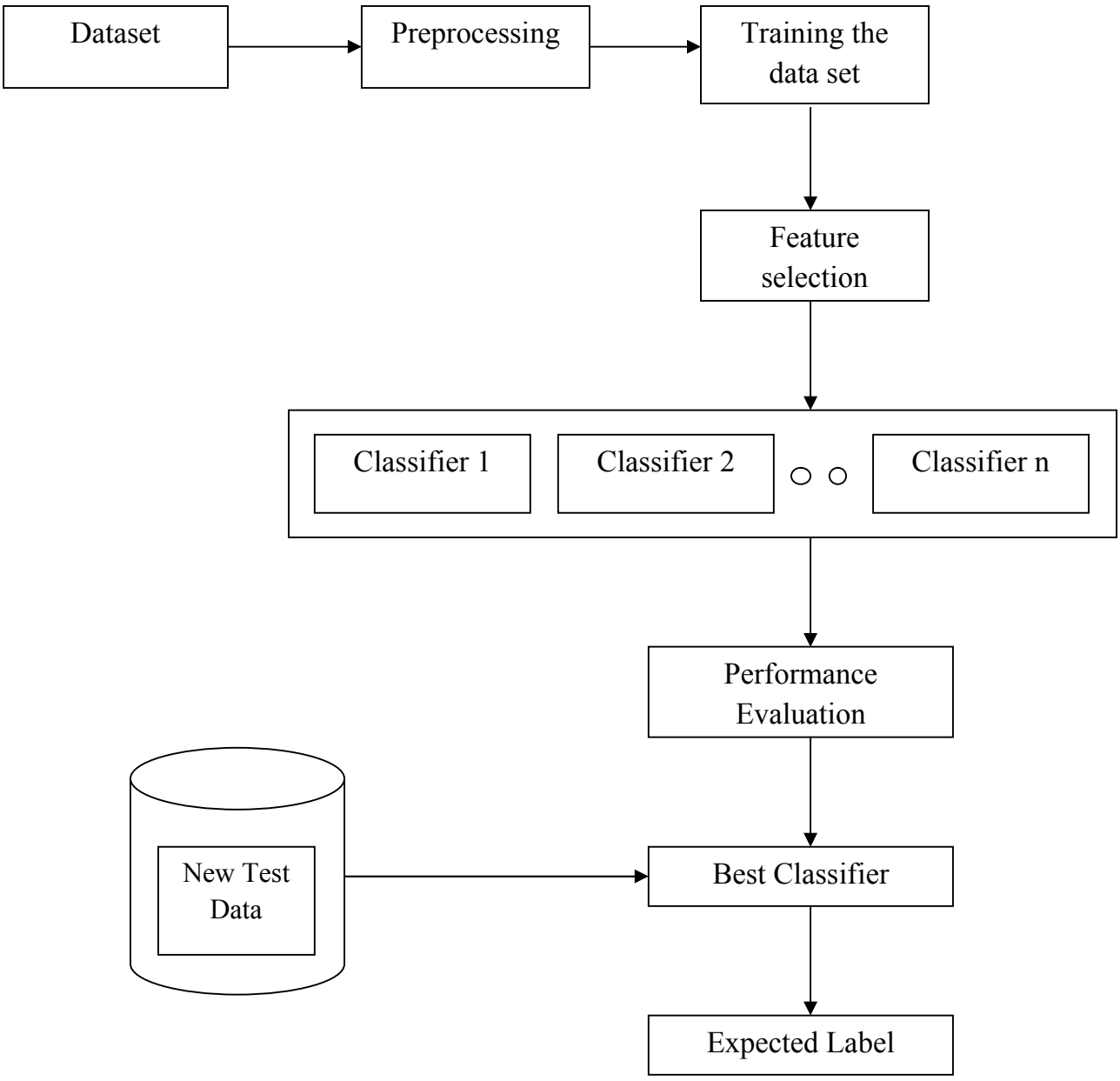
Performance

The system is developed in high level language so it will give response in very less time.

Supportability

The system is designed to be cross platform supportable.

OVERALL ARCHITECTURE



PROPOSED ALGORITHM

Firstly we will reduce the dataset so as to increase the runtime of the algorithms. For that we will use stemming and filtering of different dataset which we will use. For increasing the accuracy of the project we will run our training and testing dataset over different classifiers resulting in increasing the accuracy of the project. A comparative study with the different algorithm will help us decide the best algorithm to use on the user data. So the first algorithm which we will propose is Naïve Bayes. It typically use bag of words features to identify spam, an approach commonly used in text classification. Particular words have particular probabilities of occurring in spam data and in legitimate data. The filter doesn't know these probabilities in advance and hence must be trained to build a bag of words. To train the data, the user must know beforehand which mail is spam and which is ham. After training, the word probabilities are used to compute the probability that a data with a particular set of words in it belong to either category. Each word in the data contributes to spam probability, or only the most interesting words. This contribution is called the posterior probability and is computed using Bayes' theorem. The result will then be tested on the testing dataset and we will find the probability of success from the Naïve bayes classifier approach. The other classifier we will use would be Decision Tree. In decision tree, firstly after the preprocessing of the dataset, we will select relevant attributes over which we will create our decision tree. The attributes selected should be such as to give better result. After attribute selection we will feed the training dataset to Weka j48 tool which will create our decision

tree on the basis of important attributes first and further classification based on the other attributes. The result will be fed to the testing dataset and probability will be calculated of success of detection. Different algorithm usage and the algorithm which will give us better result in specified different domains will be selected correspondingly resulting in an increase in overall accuracy. We will do a further classification of the resultant ham into promotional and primary datasets and same in spam into adult and non-adult content using different classifiers.

RISK ANALYSIS AND MITIGATION PLAN

Risk Id	Risk Description	Risk Area	Probability (P)	Impact (I)	RE= P*I	Risk Selected for Mitigation (Y/N)	Mitigation Plan	Contingency Plan
1.	Low Accuracy	Algorithm	M(3)	H(5)	3*5= 15	Y	Use different algorithms simultaneously	Not required
2.	Too much time in identifying spam	Algorithm	M(3)	H(5)	3*5= 15	Y	Use time efficient algorithms wherever possible	Not required

TEST PLAN

Testing process starts with a test plan. This plan identifies all the testing related activities that must be performed and specifies the schedules, allocates the resources, and specified guidelines for testing. During the testing of the unit the specified test cases are executed and the actual result compared with expected output.

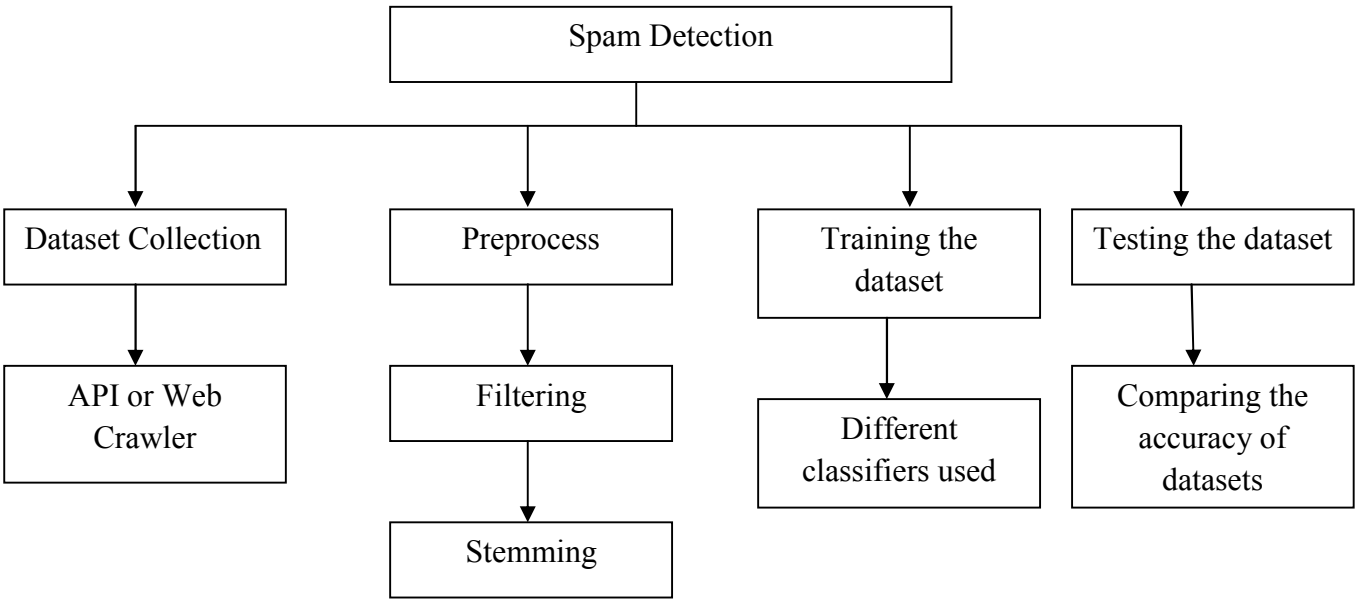
Test Data

The whole dataset is divided into some percentage into training and testing data. In the Test dataset, we have both spam and ham datasets along with their class labels. For ex. we know beforehand which dataset is ham and which is spam.

Unit Testing

The resultant from the training dataset is applied over the test dataset. The result from the test dataset is compared with their initial labels and accuracy percentage is calculated corresponding to each algorithm.

WORK BREAKDOWN STRUCTURE



REFERENCES

[1] Biju Issac, Wendy J. Jap, "Implementing Spam Detection using Bayesian and Porter Stemmer Keyword Stripping Approaches" , TENCON 2009 - 2009 IEEE Region 10 Conference on 23-26 Jan, 2009

[2] Alex Hai Wang "DON'T FOLLOW ME *Spam Detection in Twitter"*, Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on 26-28 July, 2010

[3] M. Basavaraju, Dr. R. Prabhakar, "A Novel Method of Spam Mail Detection using Text Based Clustering Approach", International Journal of Computer Applications (0975 – 8887) Volume 5– No.4, August 2010

[4] Victor M. Prieto, Manuel Alvarez´ and Fidel Cacheda," Detecting Linkedin Spammers and its Spam Nets", *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 9, 2013*

[5] M. Soiraya, S. Thanalerdmongkol, C. Chantrapornchai, "Using a Data Mining Approach: Spam Detection on Facebook", International Journal of Computer Applications (0975 – 8887) Volume 58– No.13, November 2012

[6] Shweta Rajput, Amit Arora, "Designing Spam Model- Classification Analysis using Decision Trees" , August 2013

[7] Sarit Chakraborty, Bikromadittya Mondal, "Spam Mail Filtering Technique using Different Decision Tree Classifiers through Data Mining Approach - A Comparative Performance Analysis", June 2012

[8] Megha Rathi , Vikas Pareek, "Spam Mail Detection through Data Mining – A Comparative Performance Analysis"