Survey on Email Spam Classification using Different Classification Method

[1]Priyanka Sao,[2]Prof. Anshul Singh
[1] Research Scholar M-Tech (SE),[2] Assistant Professor
[1] Computer Science & Engineering Specialized in Software Engineering,
[1] Rungta College of engg. & tech., Bhilai, India

***Abstract*— Email spam is also known as junk e-mail or Unsolicited Bulk E-mail (UBE). This junk e-mail is a subset of electronic spam involving nearly similar messages sent to many recipients by e-mail. The spam is an undesired data or facts that a web user receives in the form of e-mail or message. Spam is swamps the internet with numerous copies of the same messages to force on the people who would not otherwise choose to receive it. The work focuses on the classification method for detecting spam. And also increases the efficiency of the spam detection system.**

***Index Terms*—E-mail spam, Classification method, Feature Extraction technique.**

I. Introduction

Over the last few decades, with universalize of the Internet, junk email has become a big problem for email user; daily an astounding amount of spam has running into users' mailboxes. Spam filtering is the problem of automatically filtering unwanted electronic mail messages. In its simple form, spam filtering can be converted as text classification task where the classes to be predicted are spam or legitimate.[1]

The SMS is a short message service which is part of our daily life. For communication purpose people generally SMS to each other. Some times for the users SMS become a headache like Promotional companies send bulk SMS to the users which they don't want. Certain criteria must be decided to identify an SMS to be spam. One of the major factors in SMS spam detection is the textual analysis of the SMS. Spam emails, confuse and irritate SMS users by wasting their precious time. Spam even provides various kinds of attacks and distributed harmful content or data such as viruses, worms, Trojan horses and other malicious code. Several technical solutions are available for dealing with these issues like commercial and open-source products.

Two types of spam filtering are:

- Based on Non-machine learning
- Based on Machine learning [2].

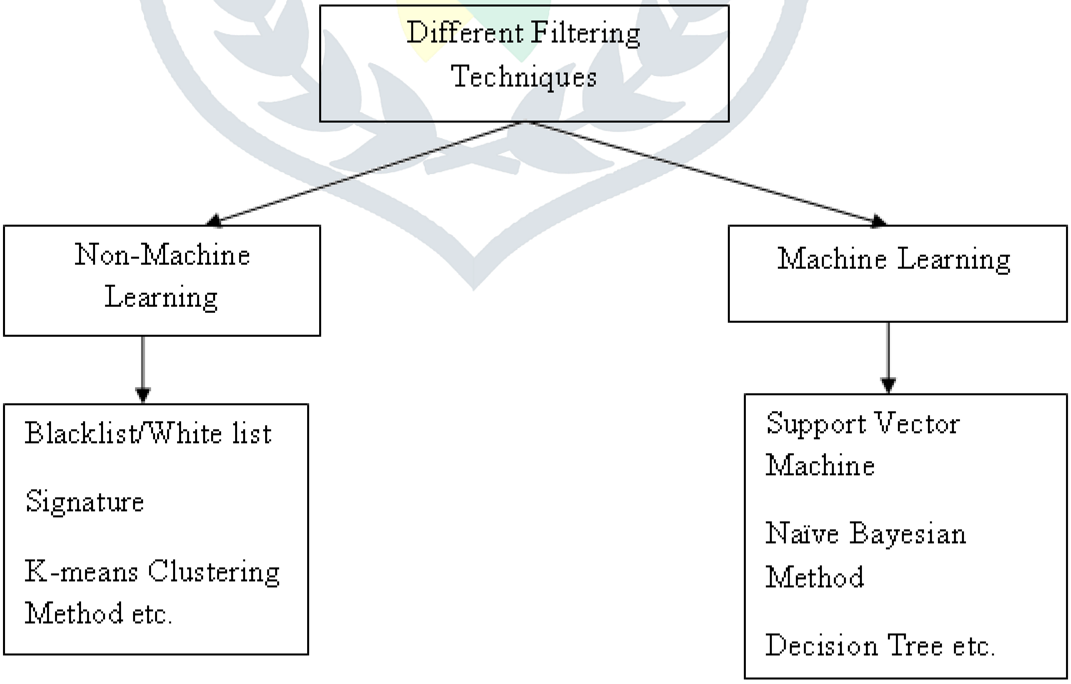

Figure 1.1: Classification of the various approaches to spam filtering

Today, information production and exchange is exposing users to a saturation of content. With automatic text categorization this challenging problem is being tackled. Automatic categorization of Arabic documents has become very important, especially that their number online is rapidly growing. One way to approach the problem of automatic text categorization is by means of supervised machine learning techniques.[3] Emails spam have several harmful effects such as downgrading users trust by advertising emails, decreasing communication quality and also companies' creditability issues. In addition, in most cases, these emails have breach of trust and fraudulent goals. The legitimate emails and also evolving nature of spam leads to change in underlying distribution of emails content.[4]

Through email, companies and individuals send advertisements for various products, undesirable harmful news, and contents, and fake proposals etc. These spam emails confuse email users and waste their precious time. In supervised or inductive machine learning, the algorithms learn from the training dataset that contains both inputs and outputs (results) and a model is created. The model is then tested for new samples for classification.[5] As the number of Internet users has increased email has become one of the most reliable and economical forms of communication. Individuals and organizations rely more and more on the emails to communicate and share information and knowledge.[6]

In today's electronics-based world, many traditional forms of communication are being replaced with email. The advantage of the emails is ease of communication with a large number of individuals. This advantage, however, is also captured by senders of unsolicited bulk emails (UBE). UBE can be divided in two related but distinct classes: Spam and Phishing. [7]

This paper present a method based on artificial immune system to detect spam message. An artificial immune system (AIS) is a computational model inspired bi mammalian immune system.[8] Spam emails are actually not just waste of resources, but also act serious security problems. The major problem here is to fast detect new burden of spam emails, where previous knowledge is not present. The spammers constantly adapt new methods in order to beating spam filters.[9]

Unsolicited Bulk Email (UBE) has become a huge problem in recent years. The effect of this on valid users is many fold.

1. Spammers often advertise products / services that may be harmful or offensive to recipients such as unlicensed medicines or pornography.
2. Recipients productivity is decreased as they read spam emails.
3. Mail server efficiency is decreased due to the extra volume of incoming email. [10]

Most of the modern e-mail software packages provide some form of programmable filtering techniques, in which the packages are in the form of rules that organize mail into files or folders of spam mail based on keywords which are detected in the header or body of mail. [11]

1.1 The evolution of spam

Until a minute past, spam was the domain of text- or html-based emails. Spammers designed personalized template emails to deliver their messages and so created use of bulk mailing software system for distribution.

Today, most of the people are believe on e-mail to connect them with their, family, friends, customers, colleagues, and also business partners. Spam is a very critical problem that is possibly threatens the presences of e-mail services. In particular, it is now a nontrivial task to find legitimate e-mails which are messy with the spam e-mails in an e-mail inbox.[12]

Two categories for Filter classification strategies: those based on machine learning (ML) principles and those not based on ML. Non-machine learning techniques, such as heuristics, blacklisting and signatures, have been complemented in recent years with new, ML-based technologies.[13] In general these are broadcast messages send to a large number of peoples. However, it is important to differentiate between solicited email and unsolicited email, which can be labeled as Spam. Spam's are not only wastage of money, bandwidth also very irritating for the users.[14]

The classification problems can be treated in business, medicine, and industry and science problem. A classification problem can be occurs when an object needs to be assigned into a predefined group or class based on a number of discovered attributes related to that object. Since any classification procedure seeks a functional relationship between the group membership and the attributes of the object, accurate identification of this underlying function is certainly important.

II. LITERATURE REVIEW

Sun X., Zhang Q. and Wang Z. (2009) introduce to comfortably deal with spam mail filtering problem, a novel spam filtering algorithm based on

locality pursuit projection (LPP) and least square version of SVM(LS-SVM) is proposed in this paper.

Sharma A. and Anchal (2014) focuses on the text classification methods like tree architecture, ICA algorithm and Neural Network algorithm for the text classification to prevent the user from unwanted spam messages or data. The main aim of this paper to examine the idea of text classification methods like fast ICA and neural networks. For pattern matching and feature extraction technique are used to increase accuracy of the system.

Belkebir R. and Guessoum A. (2013) present the three different types of approaches based on Arabic text categorization. These approaches are: artificial neural networks, support vector machine (SVMs) and a hybrid approach BSOCHI-SVM. They are explaining the approach and also present the results of the evaluation and implementation using two types of representations which are root-based stemming and light stemming.

Hayat M., Basiri J., Seyedhossein L., and Shakery A. (2010) and Azadeh Shakery introduce an adaptive spam filtering system which is based on language model. Adaptive spam filtering system proposed which can detect concept drift based on computing deviation in email contents distribution. This proposed method can also be used along with any existing classifier. In this paper use Naïve Bayes method as classifier. The results provide information of the method in detecting concept drift efficiently and also its superiority over Naïve Bayes classifier in terms of accuracy.

Panigrahi P. (2012) In this paper discusses performance of different supervised machine learning techniques such as Bayes algorithms, tree algorithms, neural network, and support vector machines for classifying a spam e-mail. And this result maintained by UCI Machine Learning Repository and also compare these methods. The result found that neural network provides the best result among all the classifiers.

Ali M., Gharan O., and Raahemifar K. (2014) present a novel algorithm to accurately streamer junk email and also to separate Spam email from Ham email. The error rate of a single optimization algorithm will improve by 39% using of our consultation and voting (CAV) algorithm. The result is the best classification algorithms are the Rnd Tree classification algorithm and the Fisher filtering feature selection algorithm. The best error rate is 0.0089.

Sharma D. et al (2009) for detecting the spam emails without using prior knowledge it has proposed the feasibility of negative selection algorithm. They use TREC07 collections for the experiments.

Toolan F. and Carthy J. (2010) in this paper present the issue of 40 feature which are used in recent literature for identify the utility of features. Information gain for these features is calculated over Ham, Spam and Phishing corpora. From this created C5.0 classifiers using three groups of features, those with the best IG values, the median IG values, and finally the worst IG values. As expected, in each case, the classifier trained on the best features outperformed all of the others.

Clark J., Koprinska I., and Poon J. (2010) present a automated e-mail filing into folders and ant spam filtering for neural network based system. For automated email filing into spam mail filtering the Neural network approach is used. For the email filing tf-idf weighting and frequency with mailbox level normalization produced the best results.

Hameed M. and Mohammed N. (2013) provide the information about automated tool to identify whether a message is spam or not based on the content of the message. For this purpose they introduce (OBP) “Optical Back Propagation” technique. In term of accuracy, precision, recall, false Positive, false negative, and speed of the net the OBP spam filtering is reasonable.

Negi S. and Negi R. (2014) discuss some approaches for spam detection. This paper explored different approaches to deal with spam problem. After examination from all of these approaches no one can provide 100% result. But some of the approaches provide high false positive rates and false negative rates.

Shrivastava J. and Maringanti H.(2014) for spam email filtering they introduce a genetic algorithm based method and also discussed with its advantages and dis-advantages. The results presented in the paper are suggested that genetic algorithm can be a good option in conjunction as compared to other e-mail filtering techniques and can provide more robust solution.

III. METHODOLOGY

Spam classification has contained the different machine learning classification, Naïve Bayesian Classifier has one of the popular email spam classification method. [4] An efficient filtering algorithm based on LPP and LS-SVM is described. The high-dimensional email messages are first mapped into lower-dimensional feature space with LPP, the LS-SVM classifier is then applied to classify the email messages into semantically different classes.[1]

In supervised learning process text classification is very popular. In supervised learning process a task is assign to the text data or document and then classifies this text data according to predefined categories or classes according to their contents. According categories of their contents the data is automatically classified. Now days there are different types of algorithm are present to deal with automatic text classification. For this purpose it uses the FastICA algorithm and Support Vector Machine classifier.[2]

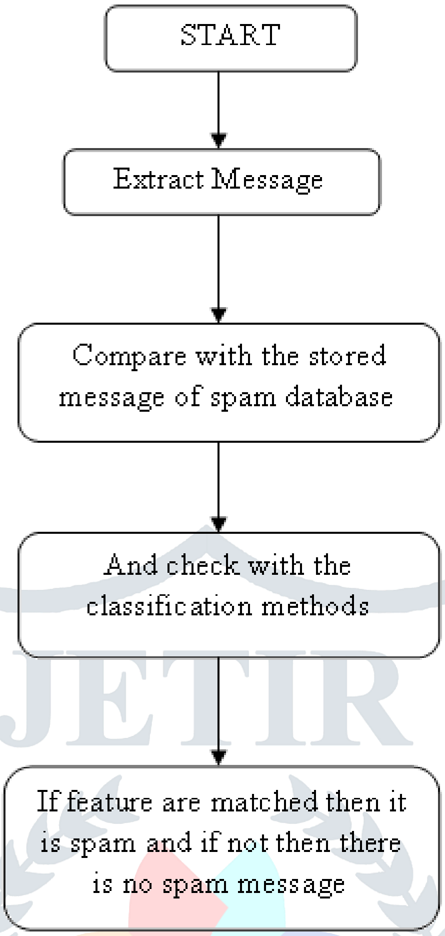
START
Extract Message
Compare with the stored message of spam database
And check with the classification methods
If feature are matched then it is spam and if not then there is no spam message

Fig 3.1: The Proposed Methodology

In this Method to explain about the technique how to identify the spam message or mails. Firstly we start to extract the data from the files or any kind of mails and compare with the spam database. The various classification techniques are present which are used to identify the spam mails. If the data are matched with the spam domain then this mail is spam mails and if not matched then it is not a spam mails.

Different Classifier Methods:
We study the different classifier method for detecting the spam mails these are:

- Support Vector Machine (SVM)
- Negative Selection Algorithm
- Naïve Bayesian Classifier
- Decision Tree
- K-means Clustering Method

Support Vector Machine: Support Vector Machine is supervised learning method. SVM is used for classification, as well as used for regression. Most important thing in Support Vector Machine is the data to be separated needs to be binary.

Negative Selection Algorithm: A Negative Selection Algorithm is a classification algorithm that simulates the process of negative selection in the creature immune system. Negative selection is a branch of AIS, simulates the way a human body detects and destroys harmful antigens.

Naïve Bayesian Classifier: A Naive Bayesian classifier is a simple probabilistic classifier which are based on a Bayesian' theorem with naïve (strong) independence assumptions. A more descriptive term for the probability model would be "independent feature model". It requires only the small amount of training data for estimation. Because independent variables are assumed, not the entire covariance matrix need to be determined and only the variances of the variables for each class are to be determined.

Decision Tree: Decision tree is used for representing the decision. Decision Tree decides the dependent variable (target value) of a new sample based on various attribute values of a decision tree which is denote the different attributes. The decision tree are represents like a flow chart where each non leaf node represent the test data, each branch are represent the result or outcomes of that test data and at the last each leaf node denotes a class label.

K-means Clustering Method: The K-means Clustering Algorithm which are used in SenseClusters. When we divide the available data instances into given number of sub-groups called clustering process. The sub-groups are called clusters, and also we can say "Clustering". In a simple way, to cluster a particular set of instances into K different clusters, where K is a positive integer this is called K-means algorithm.

IV. Conclusion

Spam is an unwanted data or facts that a web user receives in the form of email or message. Spams are the textual conditions of the system and these conditions are cause failure to our system. For this purpose Most of the current methods for spam filtering are used to separate spam and regular message. This can result to protect our system for any damage and failure. This paper focuses on some classification and filtering methods to prevent the user for unwanted spam. To study some methods for classify the spam message and according to that the Naïve Bayesian Classifier are well suited for identifying spam mails.

References

[1] Sun X., Zhang Q. and Wang Z. (2009), "Using LPP and LS-SVM For Spam Filtering", School of Information Science and Engineering Henan University of Technology IEEE 2009, pp. 4244-4246.

[2] Sharma A. and Anchal (2014), "SMS Spam Detection Using Neural Network Classifier",ISSN: 2277 128X Volume 4, Issue 6, June 2014, pp. 240-244.

[3] Belkebir R. and Guessoum A. (2013), "A Hybrid BSO-Chi2-SVM Approach to Arabic Text Categorization", IEEE 2013, pp. 978-984.

[4] Hayat M., Basiri J., Seyedhossein L., and Shakery A. (2010), "Content-Based Concept Drift Detection for Email Spam Filtering", 5th International Symposium on Telecommunications, IEEE 2010.

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

[6] Ali M., Gharan O., and Raahemifar K. (2014), , "Multiple Classifications for Detecting Spam email by Novel Consultation Algorithm", CCECE 2014, IEEE 2014, pp. 1-5.

[7] B. Sirisanyalak and O. Sornil, (2007) "An Artificial Immunity-Based Spam Detection System",IEEE 2007.

[8] Toolan F. and Carthy J. (2010), "Feature Selection for Spam and Phishing Detection", IEEE 2010, pp. 1-12.

[9] Clark J., Koprinska I., and Poon J. (2010), " A Neural Network Based Approach to Automated Email Classification.

[10] Hameed M. and Mohammed N. (2013), "A Content based Spam Filtering Using Optical Back Propagation Technique", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 7, July 2013, pp. 416-421

[11] Negi S. and Negi R. (2014), "A Review on Different Spam Detection Approaches", International Journal of Engineering Trends and Technology (IJETT) – Volume 11 Number 6 - May 2014, pp. 315-318.

[12] Shrivastava J. and Maringanti H.(2014), "E-mail Spam Filtering Using Adaptive Genetic Algorithm", I.J. Intelligent Systems and Applications, 2014, pp. 54-60.

Priyanka Sao

Mtech Scholar from Computer Science & Engineering specialised in Software Engineering from Rungta College of Engineerin & Technology Bhilai (C.G.), done BE in Information Technology from Central College of Engineerin & Management Raipur (C. G.)

Prof. Anshul Singh

Work as Assistant Professor from Computer Sceince & Engineering depatment in RCET done BE in Information Technology in 2010 from CSVTU, done M-Tech in Computer Sceince & Engineerin in 2012 from CSVTU. Teaching Experience 4 years in RCET Bhilai publish 2 national and 4 international papers.