Paper on Spam Email Detection with Classification Using Machine Learning

Naresh Vinod Wankhade[1], Dr.Ranjit. R. Keole[2], Prof.Tushar. R. Mahore[3]
[1]*ME (Computer Science and Engineering) Second Year, Dr.RGIT&R, Amravati, India*
[2]*Head of the Department, Information Technology, HVPM's CET, Amravati, India*
[3]*Head of the Department, Computer Science & Engineer DRGIT&R, Amravati, India*

Abstract— **The increasing volume of unsolicited bulk e-mail (also known as spam) has generated a need for reliable anti-spam filters. Machine learning techniques now days used to automatically filter the spam e-mail in a very successful rate. In this paper we review some of the most popular machine learning methods (Bayesian classification, k-NN, ANNs, SVMs, Artificial immune system and rough sets) and of their applicability to the problem of spam Email classification. Descriptions of the algorithms are presented, and the comparison of their performance on the Spam Assassin spam corpus is presented. Electronic mail has eased communication methods for many organizations as well as individuals. This method is exploited for fraudulent gain by spammers through sending unsolicited emails. This article aims to present a method for detection of spam emails with machine learning algorithms that are optimized with bio-inspired methods. A literature review is carried to explore the efficient methods applied on different datasets to achieve good results. Extensive research was done to implement machine learning models using Naïve Bayes, Support Vector Machine, Random Forest, Decision Tree and Multi-Layer Perceptron on seven different email datasets, along with feature extraction and pre-processing. The bio-inspired algorithms like Particle Swarm Optimization and Genetic Algorithm were implemented to optimize the performance of classifiers. Multinomial Naïve Bayes with Genetic Algorithm performed the best overall. The comparison of our results with other machine learning and bio-inspired models to show the best suitable model is also discussed.**

Index Terms- **ANN, Data Extraction, IP Filtration, Machine Learning, URL**

I. INTRODUCTION

Recently unsolicited commercial/bulk e-mail also known as spam, become a big trouble over the internet. Spam is waste of time, storage space and communication bandwidth. The problem of spam e-mail has been increasing for years. In recent statistics, 40% of all emails are spam which about 15.4 billion email per day and that cost internet users about $355 million per year. Automatic e-mail filtering seems to be the most effective method for countering spam at the moment and a tight competition between spammers and spam-filtering methods is going on. Only several years ago most of the spam could be reliably dealt with by blocking e-mails coming from certain addresses or filtering out messages with certain subject lines. Spammers began to use several tricky methods to overcome the filtering methods like using random sender addresses and/or append random characters to the beginning or the end of the message subject line [11]. Knowledge engineering and machine learning are the two general approaches used in e-mail filtering. In knowledge engineering approach a set of rules has to be specified according to which emails are categorized as spam or ham. A set of such rules should be created either by the user of the filter, or by some other authority (e.g., the software company that provides a particular rule-based spam-filtering tool). By applying this method, no promising results shows because the rules must be constantly updated and maintained, which is a waste of time, and it is not convenient for most users. Machine learning approach is more efficient than knowledge engineering approach; it does not require specifying any rules [4]. Instead, a set of training samples, these samples is a set of pre classified e-mail messages. A specific algorithm is then used to learn the classification rules from these e-mail messages. Machine learning approach has been widely studied and there are lots of algorithms can be used in e-mail filtering. They include Naïve Bayes, support vector machines, Neural

Networks, K-nearest neighbor, rough sets and the artificial immune system.

The proposed system will help to enhance the security of user through previous checking of email. In which the evolutionary mechanism firstly check the content of the mail which passed through various machine learning technique. In this the proposed methodology will perform the various check for the link as well which will help for the security enhancement. It will handle the cyber security attack to stop the entry.

2. EXISTING SYSTEM& ALGORITHM

There are some research works that apply machine learning methods in e-mail classification, Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali[2] . They demonstrated that the naïve Bayes e-mail content classification could be adapted for layer-3 processing, without the need for reassembly. Suggestions on redetecting e-mail packets on spam control middle boxes to support timely spam detection at receiving e-mail servers were presented. M. N. Marsono, M. W. El-Kharashi, and F. Gebali[1] They presented hardware architecture of na¨ıve Bayes inference engine for spam control using two class e-mail classification. That can classify more 117 million features per second given a stream of probabilities as inputs. This work can be extended to investigate proactive spam handling schemes on receiving e-mail servers and spam throttling on network gateways. Y. Tang, S. Krasser, Y. He, W. Yang, D. Alperovitch [3] proposed a system that used the SVM for classification purpose, such system extract email sender behavior data based on global sending distribution, analyze them and assign a value of trust to each IP address sending email message, the Experimental results show that the SVM classifier is effective, accurate and much faster than the Random Forests (RF) Classifier. Yoo, S., Yang, Y., Lin, F., and Moon [11] developed personalized email prioritization (PEP) method that specially focus on analysis of personal social networks to capture user groups and to obtain rich features that represent the social roles from the viewpoint of particular user, as well as they developed a supervised classification framework for modeling personal priorities over email messages, and for predicting importance levels for new messages. Guzella, Mota-Santos, J.Q. Uch, and W.M. Caminhas[4] proposed an immune-inspired model, named innate and adaptive artificial immune system (IA-AIS) and applied to the problem of identification of unsolicited bulk e-mail messages (SPAM).

3. PROPOSED METHODOLOGY

As per the things seen it is necessary to propose the mechanism in which mail are going to cross verify the mail content in which we are going to filter the both content and links of shared email. Most probably the spam mails contain the malicious link in which URL classification or parsing need to be work out. So that in proposed we analyze the URL data as well as mail content

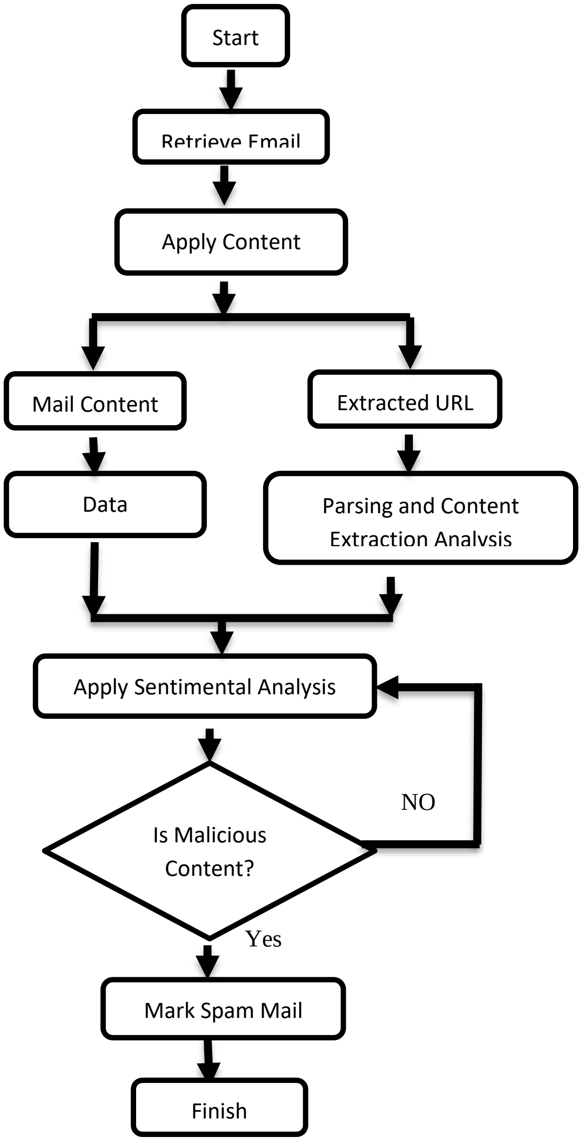

Fig 1 –Flow chart of Proposed Methodology for spam email detection Above diagram represents the flow chart of proposed methodology in which mails are

given as input to the system in which on mail content the content extraction will be done and followed with execution process of breaking it in to the links and data in this it is going to filter in various aspect like content filtration counting the malicious word and shows it in appropriate manner firstly the link and data classification will be workout latterly the data process with sentimental analysis in which the various keywords compared and evaluate . Latterly the step of IP check will be encounter in which the send email id will be retrieve and perform with evaluation. This process followed by result evaluation. At the end the spam email detection will be concluded

Architecture

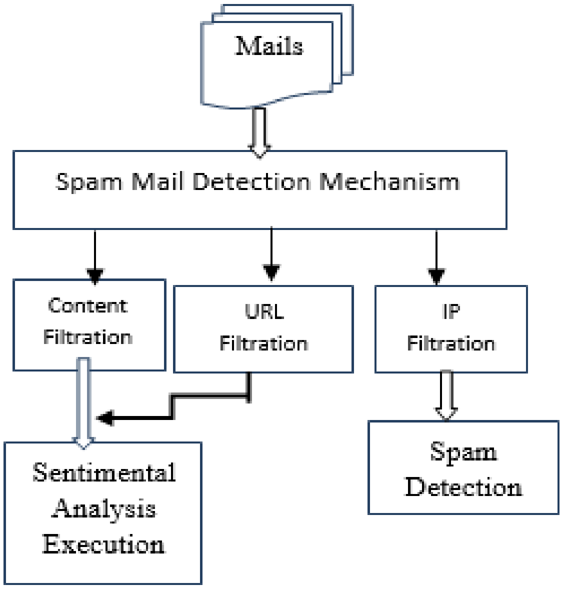

Fig 2-Execution Spam email

Above diagram represents the architecture of execution of spam email in which the first step will be perform as content filtration URL extraction and separating the data . in this the link-based evaluation well done the content of the mail will be compared with existing keyword and IPs. So that the spam email detection will be done.

Implementation

Spam email detection system has following stages in order to detect spam email. Admin has access to all contents in the Spam email detection system. New User has to register using Username, email id, Contact details and password to be chosen by the user. After registering user can access his account by login using email id as a username and password. Registered user (Sender) can send mail to another registered user (Receiver) by selecting appropriate email id. At the receiver end each mail has to go through stages discussed in system design and implementation of all stages is as follow.

Steps in Spam email Detection system

Admin Login

Admin has access to all contents in the Spam email detection system, Admin can make certain changes. Following screenshot shows the login window for admin

New Registration Window

New User has to register using Username, email id, contact details and password to be chosen by the user. User has to remember all credentials in order to access account under user login. In this case, email id shall be used as username. Screenshot for New Registration is shown below.

User Login Window

Registered user has to use emailed as username to login into the account. Once login is done, registered user (Sender) can send mail to another registered user (receiver) using appropriate email id.

Spam Email Detection System

All registered users can be accessed by admin. Admin can certainly changes to spam mail, users, can remove spam and can add training.

However, the fields are confined to username, contact number, email id which is later used as username for login and password. In this system one user can have same username with different email ids. Here in this case, email id acts as a primary key. Duplication of email id strictly restricted here in Spam Email Detection System

Spam Detection Mail Window

After login into user account, registered users can make certain changes to spam mail, users, can remove spam and can add training. However, the fields are confined to username, contact number, email id which is later used as username for login and password. In this system one user can have same username with different email ids. Here in this case, email id acts as a primary key. After receiving mail, user can check the mail body, if he finds inappropriate word then he can report that mail as spam or otherwise non spam. Under this mechanism it performs actions like Mail Splitting, Content Extraction ,URL Filtration, IP Extraction to detect the spam

Result Analysis

We pass certain email content to both existing and proposed in which all mails are pass which are non-spam so below system show the detections of mails which are shown below

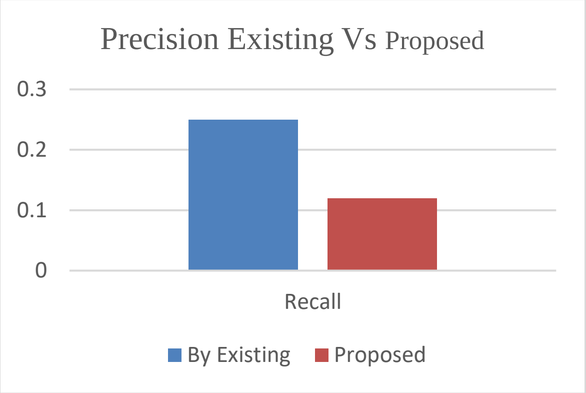

Fig.3- Precision Existing Vs Proposed System

4.CONCLUSION

In Spam mail classification is major area of concern these days as it helps in the detection of unwanted emails and threats. So now a day's most of the researchers are working in this area in order to find out the best classifier for detecting the spam mails. So a filter is required with high accuracy to filter the unwanted mails or spam mails. In this paper we focused on finding the best classifier for spam mail classification using Data Mining techniques. So, we applied various classification algorithms on the given input data set and check the results. From this study we analyze that classifier works well when we embed feature selection approach in the classification process that is the accuracy improved drastically when classifiers are applied on the reduced data set instead of the entire data set. As in proposed the spam classification done on all parameters like IP , Previous history and content of shared URL and data so that the proposed mechanism will helps a lot to go improved spam mail detection.

REFERENCE

[1] Shukor Bin AbdRazak, Ahmad Fahrulrazie Bin Mohamad "Identification of Spam Email Based on Information from Email Header" 13th International Conference on Intelligent Systems Design and Applications (ISDA), 2013.

[2] Mohammed Reza Parsei, Mohammed Salehi "E-Mail Spam Detection Based on Part of Speech Tagging" 2nd International Conference on Knowledge Based Engineering and Innovation(KBEI), 2015.

[3] Sunil B. Rathod, Tareek M. Pattewar "Content Based Spam Detection in Email using Bayesian Classifier", presented at the IEEE ICCSP 2015 conference.

[4] AakashAtulAlurkar, Sourabh Bharat Ranade, Shreeya Vijay Joshi, SiddheshSanjay Ranade, Piyush A. Sonewa, Parikshit N.Mahalle, Arvind V. Deshpande "A Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques", 2017.

[5] KritiAgarwal, Tarun Kumar "Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization", Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018.

[6] CihanVarol, HezhaM. TareqAbdulhadi "Comparison of String-Matching Algorithms on Spam Email Detection", International Congress on Big Data, Deep Learning and Fighting CyberTerrorism Dec 2018.

[7] Duan, Lixin, Dong Xu, and Ivor Wai-Hung Tsang. "Domain adaptation from multiple sources: A domain dependent regularization approach." IEEE Transactions on Neural Networks and Learning Systems 23.3 (2012).

[8] Anitha, PU &Rao, Chakunta& ,T.Sireesha. (2013). A Survey On: E-mail Spam Messages and Bayesian Approach for Spam Filtering. International Journal of Advanced Engineering and Global Technology (IJAEGT). 1. 124- 136.

[9] Attenberg, J., Weinberger, K., Dasgupta, A., Smola, A., &Zinkevich, M. (2009, July). Collaborative email-spam filtering with the hashing trick.In Proceedings of the Sixth Conference on Email and Anti-Spam.

[10] Awad, W. A., &ELseuofi, S. M. (2011). Machine learning methods for spam e-mail classification.

International Journal of Computer Science & Information Technology (IJCSIT), 3(1), 173-184.