ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

Review paper on Spam Email Detection with Classification Using Machine Learning

Naresh Vinod Wankhade[1], Ranjit. R. Keole[2], Tushar. R. Mahore[3]
Scholar, Dr.RGIE&R, Amravati, India[1]
Professor & Head of the Department, Information Technology, HVPM's CET, Amravati, India[2]
Head of the Department, Computer Science & Engineering, DRGIT&R, Amravati, India[3]

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. An 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.

Keywords: ANN, Data Extraction, URL, Machine Learning, IP Filtration

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.

Machine learning models have been utilized for multiple purposes in the field of computer science from resolving a network traffic issue to detecting a malware. Emails are used regularly by many people for communication and for socializing. Security breaches that compromises customer

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

data allows 'spammers' to spoof a compromised email address to send illegitimate (spam) emails. This is also exploited to gain unauthorized access to their device by ricking the user into clicking the spam link within the spam email that constitutes a phishing attack [1].

Many tools and techniques are offered by companies in order to detect spam emails in a network. Organizations have set up filtering mechanisms to detect unsolicited emails by setting up rules and configuring the firewall settings. Google is one of the top companies that offer 99.9% success in detecting such emails [2]. There are different areas for deploying the spam filters such as on the gate way (router), on the cloud hosted applications or on the user's computer. In order to overcome the detection problem of spam emails, methods such as content-based filtering, rule-based filtering or Bayesian filtering have been applied. Unlike the 'knowledge engineering' where spam detection rules are set up and are in constant need of manual updating thus consuming time and resources, Machine learning makes it easier because it learns to recognize the unsolicited emails (spam) and legitimate emails (ham) automatically and then applies those learned instructions to unknown incoming emails [2].

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.

II. 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 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 a secure hash message authentication code. A secure hash message authentication code to avoid certificate revocation list checking is proposed for vehicular ad hoc networks (VANETs). The group signature scheme is widely used in VANETs for secure communication, the existing systems based on group signature scheme provides verification delay in certificate revocation list checking. In order to overcome this delay this paper uses a Hash message authentication code (HMAC).

Machine Learning In E-Mail Classification

Machine learning field is a subfield from the broad field of artificial intelligence, this aims to make machines able to learn like human. Learning here means understood, observe and represent information about some statistical phenomenon. [5] discussed that the activity related status data will be communicated consistently and shared among drivers through VANETs keeping in mind the end goal to enhance driving security and solace. Finally, the e-mail classification phase of the process finds the actual mapping between training International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011 175 set and testing set. In the following section we will review some of the most popular machine learning methods.

Naïve Bayes classifier method

In 1998 the Naïve Bayes classifier was proposed for spam recognition. Bayesian classifier is working on the dependent events and the probability of an event occurring in the future that can be detected from the previous occurring of the same event [12]. This technique can be used

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

to classify spam e-mails; words probabilities play the main rule here. If some words occur often in spam but not in ham, then this incoming e-mail is probably spam. Naïve bayes classifier technique has become a very popular method in mail filtering software. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham email in its database. If the total of words probabilities exceeds a certain limit, the filter will mark the e-mail to either category. Here, only two categories are necessary: spam or ham. Almost all the statistic-based spam filters use Bayesian probability calculation to combine individual token's statistics to an overall score [1], and make filtering decision based on the score. The statistic we are mostly interested for a token T is its spamminess (spam rating) [10],

K-nearest neighbor classifier method

The k-nearest neighbour (K-NN) classifier is considered an example-based classifier, that means that the training documents are used for comparison rather than an explicit category representation, such as the category profiles used by other classifiers. As such, there is no real training phase. When a new document needs to be categorized, the k most similar documents (neighbours) are found and if a large enough proportion of them have been assigned to a certain category, the new document is also assigned to this category, otherwise not . Additionally, finding the nearest neighbours can be quickened using traditional indexing methods. To decide whether a message is spam or ham, we look at the class of the messages that are closest to it. The comparison between the vectors is a real time process. This is the idea of the k nearest neighbor algorithm:

Stage1. Training Store the training messages.

Stage2. Filtering Given a message x, determine its k nearest neighbours among the messages in the training set. If there are more spam's among these neighbours, classify given message as spam. Otherwise classify it as ham. The use here of an indexing method in order to reduce the time of comparisons which leads to an update of the sample with a complexity O(m), where m is the sample size. As all of the training examples are stored in memory, this technique is also referred to as a memory-based classifier [6]. Another problem of the presented algorithm is that there seems to be no parameter that we could tune to reduce the number of false positives. This problem is easily solved by changing the classification rule to the following l/k-rule:

If l or more messages among the k nearest neighbours of x are spam, classify x as spam, otherwise classify it as legitimate mail.

The k nearest neighbour rule has found wide use in general classification tasks. It is also one of the few universally consistent classification rules.

Artificial Neural Networks classifier method

An artificial neural network (ANN), also called simply a "Neural Network" (NN), is a computational model based on biological neural networks. It consists of an interconnected collection of artificial neurons. An artificial neural network is an adaptive system that changes its structure based on information that flows through the artificial network during a learning phase. The ANN is based on the principle of learning by example. There are, however the two classical kind of the neural networks, perceptron and the multilayer perceptron. Here we will International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011 177 focus on the perceptron algorithm. The idea of the perceptron is to find a linear function of the feature vector $f(x) = wT\ x + b$ such that $f(x) > 0$ for vectors of one class [2], and $f(x) < 0$ for vectors of other class. Here w = (w1 w2,…wm) is the vector of coefficients (weights) of the function, and b is the so-called bias. If we denote the classes by numbers +1 and -1, we can state that we search for a decision function d(x) = sign (w T x + b). The perceptron learning is done with an iterative algorithm. It starts with arbitrarily chosen parameters (w0,b0) of the decision and updates them iteratively. On the n-th iteration of the algorithm a training sample (x,c) is chosen such that the current decision function does not classify it correctly (i.e. sign (wn x + bn) ≠ c). The parameters (wn,bn) are then updated using the rule:

wn+1 = wn + cx bn+1 = bn + c The algorithm stops when a decision function is found that correctly classifies all the training samples.

The above description is used in the following algorithm [8].

Stage1. Training Initialize w and b (to random values or to 0). Find a training example (x,c) for which sign (w T x + b). If there is no such example, then training is completed Store the final w and stop. Otherwise go to next step Update (w,b): w := w + cx, b := b + c. Go to previous step.

Stage2. Filtering Given a message x, determine its class as sign (w T x + b)

Support Vector Machines classifier method

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships, the SVM modeling algorithm finds an optimal hyper plane with the maximal margin to separate two classes, which requires solving the

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

following optimization problem. Maximize Subject to n n Σ αi - ½ Σ αiαjyiyj K(xi , xj) i=1 i,j=1 n Σ αiyi = 0 i=1 where 0 ≤ αi ≤ b, i = 1, 2,….n Figure 1 An SVM separating black and white points in 3 dimensions T International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011 178 Where αi is the weight of training sample x1. If αi > 0, x1 is called a support vector b is a regulation parameter used to trade-off the training accuracy and the model complexity so that a superior generalization capability can be achieved. K is a kernel function, which is used to measure the similarity between two samples. A popular radial basis function (RBF) kernel functions. After the weights are determined, a test sample x is classified by Sign (a) = { To determine the values of < γ, b >, a cross validation process is usually conducted on the training dataset. [7] discussed because of various appealing focal points, agreeable correspondences have been broadly viewed as one of the promising systems to enhance throughput and scope execution in remote interchanges. The hand-off hub (RN) assumes a key part in helpful interchanges, and RN determination may considerably influence the execution pick up in a system with agreeable media get to control (MAC). If the training dataset is large, a small subset can be used for cross validation to decrease computing costs. The following algorithm can be used in the classification process. Input : sample x to classify training set T = {(x1,y1),(x2,y2),……(xn,yn)}; number of nearest neighbours k. Output: decision yp ∈ {-1,1} Find k sample (xi ,yi) with minimal values of K(xi ,xi) – 2 * K(xi ,x) Train an SVM model on the k selected samples Classify x using this model, get the result yp

Problem Statement

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. 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. In this these may fails sometimes so that there is necessary to increase the rate of spam detection

Proposed Work

As per the things seen it is necessary to proposed 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

Motivation

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. [9] discussed about diabetic retinopathy from retinal pictures utilizing cooperation and information on state of the art sign dealing with and picture preparing. The Pre-Processing stage remedies the lopsided lighting in fundus pictures and furthermore kills the fight in the picture

Objective

1. To develop an spam email filter mechanism
2. Email detection mechanism with content extraction
3. To implement URL verification model for execution
4. Spammer URL detection implementation

Proposed Methodology (FLOW CHART)

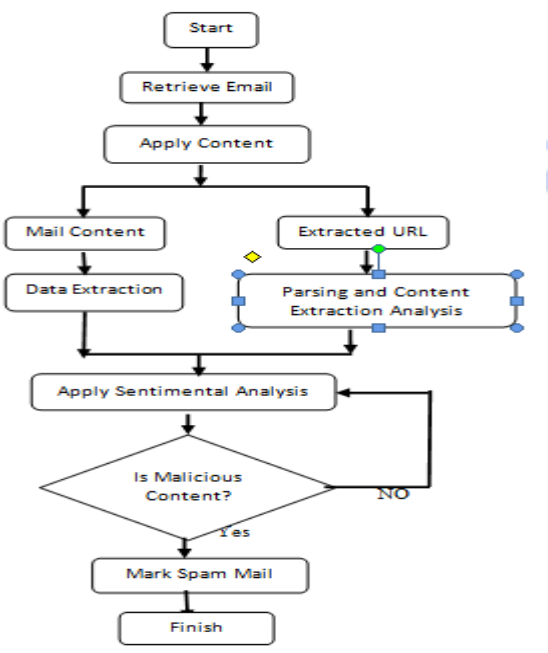

Fig 1- Flowchart of Proposed Methodology

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

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

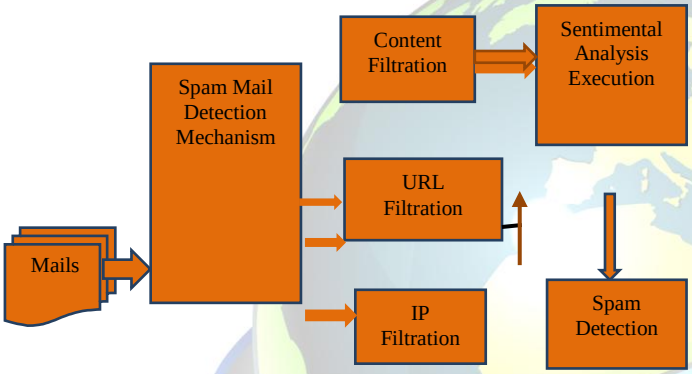

Fig 2-Execution Spam email

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.

III. 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 classifiers 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.

ACKNOWLEDGEMENT

We are grateful to the publishers mentioned in the references for providing good inputs to our review paper.

REFERENCES

1. M. N. Marsono, M. W. El-Kharashi, and F. Gebali *"Distributed Layer-3 E-Mail Classification for Spam Control"* in Proceedings of the Canadian Conference on Electrical and Computer Engineering, CCECE 2006, May 7-10, 2006, Ottawa Congress Centre, Ottawa, Canada
2. 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.
3. Christo Ananth, M.Danya Priyadharshini, "A Secure Hash Message Authentication Code to avoid Certificate Revocation list Checking in Vehicular Adhoc networks", International Journal of Applied Engineering Research (IJAER), Volume 10, Special Issue 2, 2015,(1250-1254).
4. 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.
5. Christo Ananth, Dr.S. Selvakani, K. Vasumathi, "An Efficient Privacy Preservation in Vehicular Communications Using EC-Based Chameleon Hashing", Journal of Advanced Research in Dynamical and Control Systems, 15-Special Issue, December 2017,pp: 787-792.
6. Chang, M. W., Yih, W. T., & Meek, C. (2008, August). Partitioned logistic regression for spam filtering. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 97- 105). ACM.
7. Christo Ananth, Dr. G. Arul Dalton, Dr.S.Selvakani, "An Efficient Cooperative Media Access Control Based Relay Node Selection In Wireless Networks", International Journal of Pure and Applied Mathematics, Volume 118, No. 5, 2018,(659-668).
8. [8] Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 23.
9. Christo Ananth, D.R. Denslin Brabin, Jenifer Darling Rosita, "A Deep Learning Approach To Evaluation Of Augmented Evidence Of Diabetic Retinopathy", Turkish Journal of Physiotherapy and Rehabilitation, Volume 32,Issue 3, December 2021,pp. 11813-11817.
10. Fishkin, R. (2015, November 06). Spam Score: Moz's New Metric to Measure Penalization Risk. Retrieved from https://moz.com/blog/spam-score-mozsnew-metric-to-measure-penalization-risk

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 9, Issue 8, August 2022

11. Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206- 10222.

12. Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.