JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Proposed Paper on Spam Email Detection using Machine Learning

Miss. Pratiksha Mantri[1*], Dr. Ranjit keole[2]

Student, Department of Computer Science &Engineering, HVPM COET, Amravati, India[1]
Professor &Head of Department, Department of Information Technology & Engineering, HVPM COET, Amravati, India[2]

Email id: ranjitkeole@gmail.com, pratikshamantri12@gmail.com

I. ABSTRACT

Email Spam has become a significant problem nowadays, with rapid climb of internet users, Email spams is additionally increasing. Machine learning methods are commonly utilized in spam filtering. People are using them for unethical conducts, phishing, and fraud. Sending malicious link through spam emails which might harm our system and may also seek in into your system. Creating a fake profile and email account is far easy for the spammers, they pretend sort of a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds. Machine learning techniques now days accustomed automatically filter the spam e-mail in an exceedingly very successful rate. So, it's needed to spot those spam mails which are unethical and illegal, this project will identify those spam by using techniques of machine learning, this paper will discuss the machine learning algorithms and Bio-Inspired Methods and apply all these algorithms on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy. The proposed work shows differentiating features of the content of documents [4]. There has been a lot of work that has been performed in spam filtering which is limited to some domains.

Index Terms–Machine Learning, Spam Email, Bio-Inspired, Spammers, Methodology, Spam Detection Architecture, Steaming, Spam filtering.

II. INTRODUCTION

Machine learning models are utilized for several more purposes within the field of computing from resolving a network traffic issue to detecting a malware. Emails are used daily by many people for communication and for social communication. Security contravention that compromises customer data allows 'spammers' to spoof a compromised email address to send illegitimate (spam) emails. This is also making use of unauthorized access to their device by tricking the user into clicking the spam link within the spam email, that constitutes a phishing attack [1].

Feature extraction and selection plays a vital role in the classification. In spam mail detection, email data is collected through the dataset. To obtain the accurate results, data needs to be pre-processed by removing stop words and word tokenization. Pre-processing of data is done by using TF-IDF Vectorizer module. SVM algorithm is used to detect the given email is spam or ham.

Many tools and techniques are offered by companies 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 great top companies that offers success in detecting such emails [2]. Many spam detection techniques are getting used now-a-days. The methods use filters which may prevent emails from causing any harm to the user. The contributions and their weakness are identified. There are several methods that are accessible to spam, for example location of sender, Content Based Filtering Technique is usually used to create automatic filtering rules and to classify emails using machine learning approaches, like Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbor, Neural Networks. This method normally analyses words, the occurrence, and distributions of words and phrases within the content of emails and used then use generated rules to filter the incoming email spams [28]., checking IP address or space names. [26]. Spammers or frauds use refined variations to avoid spam identification. Few measures connected with spam identification are Machine learning approaches, Naïve Bayes, Support Vector Machine, Neural Network Classification. [27] and Phishing URL is also checked by using URL filtration Technique.

The proposed spam detection to identify the difficulty of the spam classification problem are often further experimented by feature selection or automated parameter selection for the models. This research conducts various experiments involving five different machine learning models with Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). This will be compared with the bottom models to conclude whether the proposed models have improved the performance or not.

The proposed system will help to reinforce the safety of user through previous checking of email. In which the evolutionary mechanism firstly checks the content of the mail which skilled various machine learning technique. In this the proposed methodology will perform the varied check for the link also which can help for the safety enhancement. It will handle the cyber security attack to prevent the entry.

III. RELETED WORK

1.Machine Learning

Experimenters have taken a cause apply machine literacy models to detect spam emails. In the paper (3), the authors have conducted trials with six different machine learning algorithms Naïve Bayes (NB) bracket, K-Nearest Neighbor (K-NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Artificial Immune System and Rough Sets. Their end of the trial was to imitate the detecting and recognizing capability of humans. Tokenization was explored and the conception handed two stages Training and Filtering. Their algorithm sorted of four-way Dispatcher-Processing, Description of the point, Spam Bracket and Performance Evaluation. It concluded that the Naïve Bayes handed the veritably stylish delicacy, perfection, and recall. Feng etal. (1) describes a mongrel system between two machine literacy algorithms i.e., SVMNB. Their proposed system is to use the SVM algorithm and induce the hyperplane between the given confines and reduce the training set by barring datapoints. This set will also be enforced with NB algorithm to prognosticate the probability of the result. This trial was conducted on Chinese textbook corpus. They successfully enforced their proposed algorithm and there was a rise in delicacy in comparison to NB and SVM on their own. Mohammed etal. (4) aimed to detect the unsought emails by experimenting with different classifiers similar as NB, SVM, KNN, Tree and Rule grounded algorithms. They generated a vocabulary of Spam and Ham emails which is also habit to sludge through the training and testing data. Their trial was conducted with Python programing language on Dispatch-1431 dataset. They concluded that NB was the simplest working classifier followed by Support Vector Machine. Wijaya and Bisri (5) propose a mongrel- grounded algorithm, which is integrating Decision Tree with Logistic Retrogression along with False Negative threshold. They were successful in adding the performance of DT. The results were compared with the previous exploration. The trial was conducted on the Spam Base dataset. The proposed system presented as 91.67 delicacy.

2.Bio-Inspired Methods:

Agarwal and Kumar (6) experimented with NB along with Particle Mass optimization (PSO) approach. The paper applied the emails from Ling-Spam corpus and aimed to develop an enhancement in F1- score, Precision, Recall and Accuracy. The paper used Correlation Point Selection (CFS) to choose applicable features from the dataset. The dataset was separate into 6040 rates. Particle Mass Optimization was co-opted along with Naïve Bayes.

They concluded a success when their proposed combined system increased the accurateness of the detection compared to NB alone (6). Belkebir and Guessoum (7) checked the SVM algorithm along with Bee Swarm Optimization (BSO) and Chi-Squared on Arabic Text. Since there have been abundance of investigation conducted for textbook mining on English and some European languages, the authors considered to review the algorithms work on Arabic language. They experimented with three different approaches to categories automatic textbook – Neural networks, Support Vector Machine (SVM) and SVM optimizing with Bee Swarm Algorithm (BSO) along with Chi-Squared. Bee Swarming Optimization algorithm is inspired by the actions of mass of notions to achieve global result. A search area is divided and each area within the split section is assigned to other notions to explore. Every result is distributed amongst the notions and the trim result is accepted and the process is repeated until the result meets the criteria of the problem. The main problem announced is “The problem of opting the set of attributes is NP-hard”. The exploration explains the problem dealing with the point selection due to the calculation time. A vocabulary is generated and fed into the Chi2-BSO algorithm to acquire the features and eventually the achieved result is loaded within the SVM algorithm. The experimentation was carried on OSAC dataset which included textbook records. The study aimlessly named 100 textbooks from each order distributed by 7030 rates. The program performed scrapping of integers, Latin elements, cut off letters, punctuation marks and stop words. The document representation step was conducted with different modes for all approaches – SVM, BSO- Ki-SVM and artificial neural network (ANN). The SVM outperformed the ANN accomplishment time. The proposed algorithm BSO- Ki-SVM exceeds the literacy time, but it’s still linked as effective (7). The paper concluded that the proposed algorithm provides an exactness rate of95.67. They’ve also stated that SVM approach outperformed ANN. Another development is to estimate the approach of this composition on other datasets and use modes similar as n-gram or notion representation. Multiple experimenters have also probed the natural evolutionary processes to optimize the ML algorithm ‟ s performance. Taloba and Ismail (8) explored inherited Algorithm (GA) optimization by integrating it with Decision Tree (DT). The authors honor the overfitting problem with dimension of point space and attempt to overcome this issue by note line with Principal Component Analysis (PCA). The paper provides an explosive background of algorithms used and proceeds with proposed algorithm. Their program performs-processing, point importing and note line. The proposed algorithm is to find the optimal value of the parameter handed for the Decision tree (DT) algorithm. The DT algorithm used is J-48 to induce the rules and apply GA with fitness function to gain the delicacy. The program uses the BLX-α for fitness hunt and performance. Their fitness function was conducted on each existent of GA. The trial was conducted with the Enron spam dataset. The paper concluded that the GADT proposed algorithm handed advanced delicacy when compared with other classifiers without PCA. Another trial compared the performance dimension with using the PCA which handed advanced delicacy than GADT itself. Karthika and Visalakshi (9) reviews the ML algorithm – SVM along with the optimization approach – Ant Colony Optimization (ACO). The proposed algorithm was performed on the Spam Base dataset with supervised literacy system.

The paper briefly defines the being work hung on pheromone updating and fitness function. The paper provides an overview of the ML algorithm similar as NB, SVM and KNN classifiers. The proposed algorithm was conducted by co-opting the ACO algorithm into the SVM ML algorithm. ACO is hung on the actions of the ants observed while creating a shortest path towards the food source. The paper states that the proposed ACO grounded point selection algorithm deducts the memory demand along with the computational time. The trial uses N-fold cross

confirmation approach to estimate the datasets with different measures. The point selection styles were used with the ACO. The result of the proposed algorithm ACO-SVM was advanced than the rest of the ML algorithms itself. The paper concluded that the delicacy of ACO-SVM was 4 advanced than the SVM itself alone. The paper estimated that the optimization algorithm resolves the conditioning of the problem concurrently to classify the emails into ham and spam (9). Further exploration looked at algorithms for optimization similar as Firefly and Ditz hunt. The Firefly algorithm in the paper (10) was used with SVM. The experimenters experimented with the Arabic textbook with point selection. The paper concluded that the proposed system outperforms the SVM itself.

The paper (11) proposes Enhanced Ditz Hunt (ECS) for bloom sludge optimization. This is where the load of the spam term is considered. It was concluded that their proposed optimization fashion of ECS outperforms the normal Ditz hunt. The work in the below exploration has handed an insight into cross systems as well as optimization ways. The bio-inspired ways show further promising results in terms of directly detecting a spam junk mail.

IV. PROPOSED WORK

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 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. This research will experiment Bio-inspired algorithms alongside Machine learning models. This will be conducted on different spam email corpora that are publicly available. The paper aims to realize the subsequent objectives:

1) To investigate machine learning algorithms for the spam detection problem.

2) to research the workings of the algorithms with the acquired datasets.

3) To implement the bio-inspired algorithms.

4) to see and compare the accuracy of base models with bio-inspired implementation.

5) To implement the framework using Python.

Scikit-Learn library will be instigated to perform the experiments with Python, and this will enable to edit the models, conduct pre-processing, and calculate the results. The program scripts will be implemented further with the optimization techniques and compared with the base results i.e., with default parameters.

Modules of Proposed System:

The necessary stages that must be observed:

ADDING CORPUS: This section will load all the email datasets within the program and distribute into training and testing data. This process will be accepting the datasets in'*.txt' format for all email (Ham and Spam). This is to help understand the real-world issues and how can they be tackled.

Pre-processing: this is often the primary stage that's executed whenever an incoming mail is received. This step consists of tokenization.

Tokenization: this is often a process that removes the words within the body of an email. It also translates a message to its meaningful parts. It takes the e-mail and divides it into a sequence of representative symbols called tokens. In a tokenization phase every word is assigned a singular token.

Stemming: subsequent step to be performed is stemming. Stemming is employed to seek out a root of a word and thus replacing all words to their stem which reduces the number of words to be considered for representing a document. Example: sings, singing, sing have sing as their stem. In the project, we use JAVA implementation of Porter stemmer which is slightly modified to satisfy our needs. The files are named with an extension 'words stemmed'.

Stop Words: This was wont to remove the unnecessary words and characters within each email and creates a bag of words for the algorithms to match against.

Feature selection: Sequel to the pre-processing stage is that the feature selection phase. Feature selection a sort of reduction within the measure of spatial coverage that effectively exemplifies fascinating fragments of email message as a compressed feature vector. The technique is useful when the dimension of the message is large, and a condensed feature representation is required to form the task of text or image matching snappy.

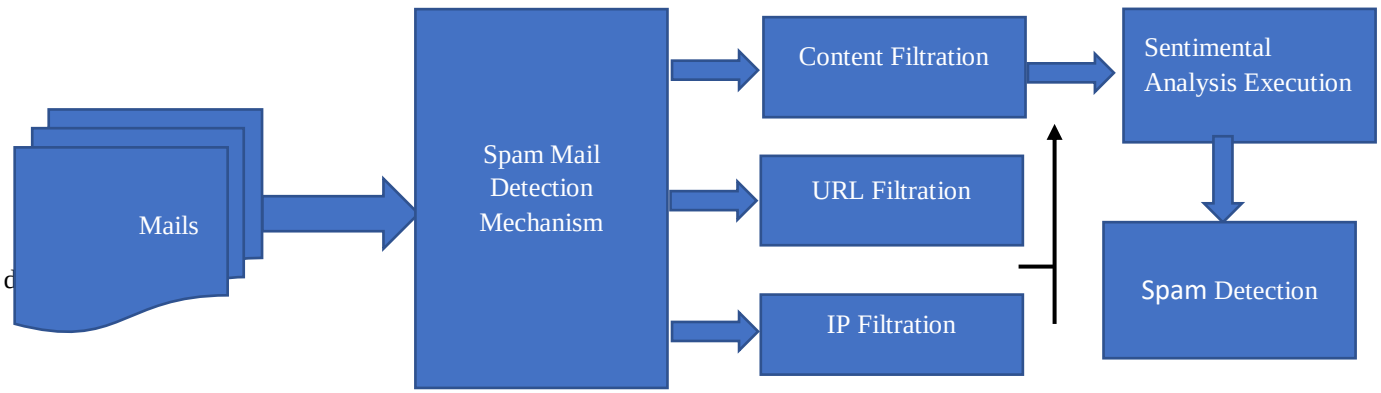
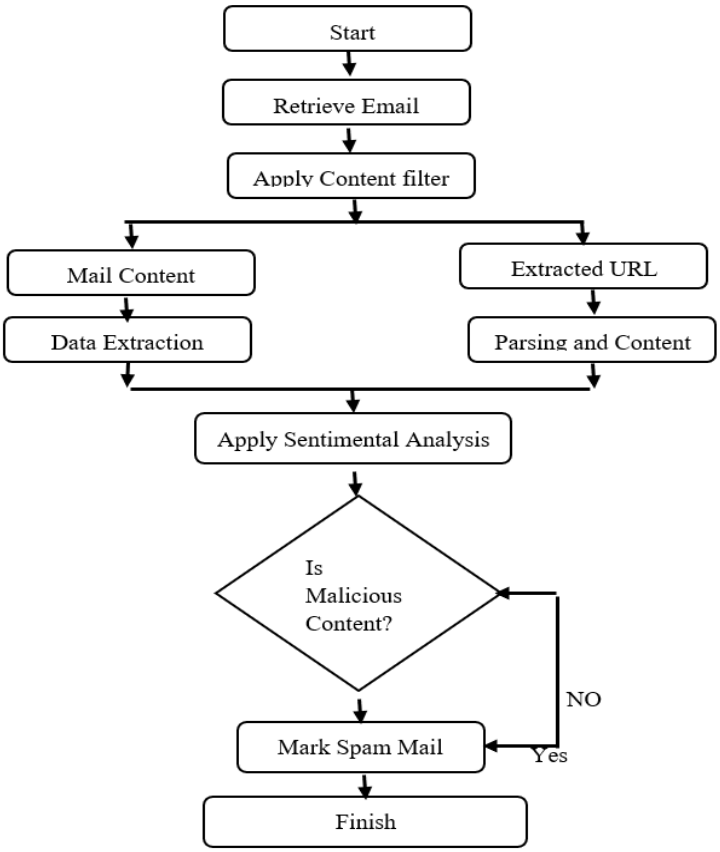


Figure 1: Spam Detection Architecture

Figure 1 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 a well done the content of the mail will be compared with existing keyword and IPs. So that the spam email detection will be done.

Email spam filtering architecture Spam filtering is aimed toward reducing to the barest minimum the number of unsought emails. Email filtering is that the process of emails to set up it in accordance with some definite standards. Mail filters are usually wanting to manage incoming mails, filter spam emails, notice and eliminate mails that contain any malicious codes like virus, trojan or malware. The workings of email are influence by some basic protocols that embody the SMTP. several the wide used Mail User Agents (MUAs) are cur, Elm, Eudora, Microsoft Outlook, Pine, Mozilla spirit, IBM notes, Kmail, and Balsa. they're email purchasers that assists the user to scan and compose emails. Spam filters is deployed at strategic places in each purchaser and servers. Spam filters are deployed by several net Service suppliers (ISPs) at each layer of the network, before of email server or at mail relay wherever there's the presence of firewall [5]. The firewall could be a network security system that monitors and manages the incoming and outgoing network traffic supported planned security rules. the e-mail server is associate degree incorporated anti-spam and anti-virus answer providing a comprehensive preventive measure for email at the network perimeter [6]. Filters is enforced in purchasers, wherever they will be mounted as add-ons in computers to function intermediator between some end devices [7]. Filters block unsought or suspicious emails that are a threat to the protection of network from going to the pc system. Also, at the e-mail level, the user will have a tailored spam filter which will block spam emails in accordance with some set conditions [8].

Proposed Methodology (Flow Chart)



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.

V. CONCLUSION

In this paper we have proposed procedure to identify an email as spam or ham based on text categorization. Different methods for pre-processing of email organize are connected, for example, applying stop words expelling, stemming, include decrease and highlight choice strategies to bring the catchphrases from every one of the qualities lastly utilizing distinctive classifiers to isolate mail as spam or ham. we reviewed machine learning approaches and their application to the area of spam filtering. The attempts made by different researchers to solving the problem of spam using machine learning classifiers was discussed. The evolution of spam messages over the years to evade filters was look out. The basic architecture of email spam filter and the

processes involved in filtering spam emails were investigated. Having mentioned the open issues in spam filtering, more analysis to reinforce the effectiveness of spam filters got to be done. this may create the event of spam filters to still be an energetic analysis field for academician and business practitioners researching machine learning techniques for effective spam filtering. Our hope is that analysis students can use this paper as a springboard for doing qualitative analysis in spam filtering mistreatment machine learning algorithms.

VI. REFERENCES

1) Deepika Mallampati, K.Chandra Shekar and K.Ravikanth "Supervised Machine Learning Classifier for Email Spam Filtering", © Springer Nature Singapore Pte Ltd. 2019 and Engineering, https://doi.org/10.1007/978- 981-13-7082-341.

2) GUPTA, H., JAMAL, M. S., MADISETTY, S., & DESARKAR, M. S. (2018, January). "A framework for real-time spam detection in Outlook." In Communication Systems & Networks (COMSNETS), 2018 10th International Conference on (pp. 380-383)

3) N. Kumar and S. Sonowal, "Email spam detection using machine learning algorithms," in Proceedings of the 2020 Second International Conference on ingenious analysis in Computing Applications (ICIRCA), pp. 108–113, Coimbatore, India, 2020

4) Veena H Bhat, Vandana R Malkani, PD Shenoy, KR Venugopal, and LMPatnaik. Classification of email using beaks: Behavior and keyword stemming. In TENCON 2011-2011 IEEE Region 10 Conference, pages1139−1143. IEEE, 2011.

5) T.S. Guzella, W.M. Caminhas, A review of machine learning approaches to spam filtering methods, Expert Syst. Appl. 36 (7) (2009) 10206–10222.

6) C.P. Lueg, from spam filtering to information retrieval and back: seeking conceptual foundations for spam filtering, Proc. Assoc. Inf. Sci. Technol. 42 (1) (2005).

7) X.L. Wang, learning to classify email: a survey, in: 2005 International Conference on Machine Learning and Cybernetics (Vol. 9, pp. 5716-5719), IEEE, Aug 2005.

8) W. Li, N. Zhong, Y. Yao, J. Liu, C. Liu, Spam filtering and email-mediated applications, in: Paper presented at the International Workshop on Web Intelligence Meets Brain Informatics, 2006.

9) G.V. Cormack, Email spam filtering: a systematic review, Found. Trends Inf. Retr. 1 (4) (2008) 335–455.

10) (2019). 3.3. Metrics and Scoring: Quantifying the standard Of Predictions— Scikit-Learn 0.22.2 Documentation. Accessed: Dec. 31, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html

11) J. Lester. (2017). Welcome to PySwarms's Documentation! —Swarms 1.1.0 Documentation. Accessed: Jan. 16, 2020. [Online]. Available: https://pyswarms.readthedocs.io/en/latest/index.html

12) R. Olson. (2019). Home—TPOT. Accessed: Jan. 12, 2020. [Online]. Available: https://epistasislab.github.io/tpot/

13) R. Shams and R. E. Mercer, ''Classifying spam emails using victimization text and readability features,'' in Proc. IEEE 13th Int. Conf. Data Mining, Dallas, TX, USA, Dec. 2013, pp. 657–666, doi: 10.1109/ICDM.2013.131.

14) T. Kumareson. (2016). Certain Investigations On Optimization Techniques to Enhance E-Mail Spam Classification. Anna University. Accessed: Feb. 26, 2020. [Online]. Available: https://shodhganga.inflibnet.ac.in/handle/10603/181292

15) H. Faris, I. Aljarah, and B. Al-Shboul, ''A hybrid approach supported particle swarm optimization and random forests for e-mail spam filtering,'' in Proc. Int. Conf. Comput. Collective Intell., 2016, pp. 498–508. [Online]. Available: https://www.researchgate.net/publication/304158714_A_Hybrid_Approach_based_on_Particle_Swarm_Optimization_and_Random_Forests_for_Email_Spam_Filtering

16) F. Temitayo, O. Stephen, and A. Abimbola, ''Hybrid GA-SVM for Efficient Feature Selection in E-mail Classification,'' Comput. Eng. Intell. Syst., vol. 3, no. 3, pp. 17–28, 2012. [Online]. Available: https://www.researchgate.net/publication/257479733_Hybrid_GASVM_for_Efficient_Feature_Selection_in_E-mail_Classification

17) Alghoul, S. Ajrami, and G. Jarousha, ''Email classification using artificial neural network,'' Int. J. Academic Eng. Res., vol. 2, no. 11, pp. 8–14, 2018. [Online]. Available: https://www.researchgate.net/publication/329307944_Email_Classification_Using_Artificial_Neural_ Network.

18) 2012 https://www.securelist.com/en/analysis/204792230/Spam_Report_April_2012

19) E.M. Bahgat, S. Rady, W. GadAn e-mail filtering approach using classification techniques The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Springer International Publishing, BeniSuef, Egypt (2016), pp. 321-331 CrossRefView Record in Scopus.