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 Review 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

The increasing volume of advice bulk e-mail (also known as spam) has generated a need for reliable anti-spam filters. Machine learning techniques now days used to directly filter the spam e-mail in a very good rate. In this paper we review some of the most desired machine learning methods (Bayesian classification, k-NN, ANNs, SVMs, Artificial immune system and rough sets) and of their bearing to the problem of spam Email classification. Electronic mail has deleted communication methods for many organizations as well as individuals. This method is used for fraudulent gain by spammers through sending unsolicited emails. This article aims to describe a method for detection of spam emails with machine learning algorithms that are optimized with bio-inspired methods. A literature review is carried to figurative the efficient methods applied on different datasets to achieve good results. This 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 evaluate the performance of classifiers. Multinomial Naïve Bayes with Genetic Algorithm performed the good overall. The comparison of our results with other machine learning and bio-inspired models to show the good suitable model is also discussed.

Index Terms–Machine Learning, Spam Email, Bio-Inspired, Spammers, Naïve Bayes Classifier Method, K-Nearest Neighbor Method, Cross-Validation.

II. INTRODUCTION

Recently unsolicited bulk e-mail also known as spam, become a big trouble over the internet. Spam is misused of time, storage space and communication bandwidth. The problem of spam e-mail has been increasing for more years. In recent statistics, 40% of all emails are spam which about 15.4 billion email each day and that cost internet users about $355 million each year. Automatic e-mail clear seems to be the most effective method for resisting spam now and a tight competition between spammers and spam-filtering methods is going on. Only several years ago most of the spam could be accurately dealt with by blocking e-mails coming from certain addresses or filtering out messages with certain subject lines. Spammers began to use several methods to overcome the filtering methods like using random sender addresses or include random characters to the beginning or the end of the message subject line [11]. methods like using random sender addresses or include random characters to the beginning or the end of the message subject line [11]. Knowledge engineering and machine learning are the two major approaches used in e-mail filtering. In knowledge engineering approach a set of rules must be specified according to which emails are assort as spam or ham. A set of rules should be created either by the user of the filter, or by some other authority by applying this method, no promising results shows because the rules must be constantly updated and maintained, which is a no use of time, and it is not convenient for most users. Machine learning approach is more structured than knowledge engineering approach; it does not require any rules [4]. Instead, a set of samples, these samples is a set of presorted e-mail messages. A specific algorithm is then used to learn the grouping rules from these e-mail messages. Machine learning approach has been extensively 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, hard sets, and the artificial immune system. Machine learning models have been utilized for many more purposes in the field of computer science 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 to gain unauthorized access to their device by tricking 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 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]. 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. To overcome the problem of spam emails, methods such as content-based formula, rule-based filtering or Bayesian filtering have been applied.

The spam detection rules are set up and are in constant need of updating thus consuming time and resources, Machine learning makes it easier because it learns to recognize the unbidden 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 checks 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.

III. LITERATURE REVIEW

There are some research work that apply machine learning technique in e-mail classification process, Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali[2] They show that the naïve Bayes e-mail content classification could be adapted for layer-3 processing, without the need of assemble again. Suggestions on predilecting e-mail packets on spam control middleboxes 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 assumption engine for spam control using two class e-mail classification.

That can classify more 117 million features per second given a stream of expectations as inputs. This work can be extended to investigate enterprising spam handling schemes on receiving e-mail servers and spam throttling on network gateways. Tang, S. Krasser, Y. He, W. Yang, D. Alperovitch [3] present a system that used the SVM for assemble purpose, such system extract email sender behavior data based on global sending distribution, analyze, and assign a value of trust to each IP address sending email message, the Experimental results show that the SVM classifier is effective, precise, and quick than the Random Forests (RF) Classifier. Yoo, S., Yang, Y., Lin, F., and Moon [11] developed personalized email prioritization (PEP) method that specially clear-cut on analysis of personal social networks to capture user groups and to obtain rich features that represent the social roles from the viewpoint of 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 congenital and flexible artificial immune system (IA-AIS) and applied to the problem of identification of unsolicited bulk e-mail messages (SPAM).It integrates entities related to macrophages, B and T lymphocytes, modeling both these changes are necessary for full immune system activation the innate and the adaptive takes place.

An implementation of the algorithm could identify more than 99% of legalized or SPAM messages in particular parameter configurations. It was compared to an upgrade version of the naive Bayes classifier, which have been attained extremely high correct classification rates. It has been finishing that IA-AIS has a greater ability to identify SPAM messages, although the identification of legitimate messages is not as high as that of the implemented naive Bayes classifier.

Existing Question Paper Generation Systems

Machine Learning in Email Classification:

Machine learning field is a compass 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 probability phenomenon. In unsupervised learning one tries to uncover hidden uniformity (clusters) or to detect anomalies in the data like spam messages or network intrusion. In e-mail clear task some features could be the bag of words or the subject line analysis. Thus, the input to e-mail classification task may be viewed as a two-dimensional matrix, whose axes are the messages and therefore the options. E-mail classification tasks are usually divided into many sub-tasks. First, information assortment and illustration are largely downside specific (i.e., e-mail messages), second, e-mail feature choice and have reduction plan to scale back the spatial property (i.e., the number of features) for the remaining steps. Finally, the e-mail classification part of the method finds the mapping between coaching International Journal of technology & data Technology (IJCSIT), Vol 3, No 1, Feb 2011 one hundred seventy-five set and testing set. within the following section we'll review several the foremost widespread machine learning strategies.

Naïve Bayes classifier method

In 1998 the Naïve Bayes classifier was projected for spam recognition. theorem classifier is functioning on the dependent events and the likelihood of an occurrence occurring within the future which will be detected from the previous occurring of a similar event [12]. this method may be accustomed classify spam e-mails; words possibilities play the most rule here. If some words occur typically in spam however not in ham, then this incoming e-mail is perhaps spam. Naïve Bayes classifier technique has become a well-liked methodology in mail filtering computer code. theorem filter ought to be trained to figure effectively. each word has sure likelihood of occurring in spam or ham email in its info. If the whole of words possibilities exceeds a definite limit, the filter can mark the e-mail to either class. Here, solely 2 classes area unit necessary: spam or ham. the majority the statistic-based spam filters use theorem likelihood calculation to mix individual token's statistics to associate degree overall score [1] and create filtering call supported the score. The data point we tend to area unit principally inquisitive about a token T is its spam (spam rating) [10]

K-nearest neighbor classifier method

The k-nearest neighbor (K-NN) classifier is considered Associate in Nursing example-based classifier, meaning that the coaching documents area unit used for comparison instead of a particular class illustration, like the class profiles utilized by different classifiers. As such, there's no real coaching section. once a brand-new document must be classified, the k most similar documents (neighbor's) area unit found and if an oversized enough proportion of them are appointed to a precise class, the new document is additionally appointed to the current class, otherwise not. in addition, finding the closest neighbors are often quickened victimization ancient compartmentalization ways. to determine

whether a message is spam or ham, we glance at the category of the messages that area unit highest to that. The comparison between the vectors could be a real time method. this is often the concept of the k nearest neighbor algorithm:
Stage1. coaching Store the coaching messages.

Stage2. Filtering Given a message x, verify its k nearest neighbors among the messages within the coaching set. If their area unit additional spams among these neighbors, classify given message as spam. Otherwise classify it as ham. the employment here of Associate in Nursing compartmentalization technique to scale back the time of comparisons that results in Associate in Nursing update of the sample with a complexness O(m), wherever m is that the sample size. As all the coaching examples area unit hold on in memory, this method is additionally named as a memory-based classifier [6]. Another downside of the bestowed formula is that there looks to be no parameter that we tend to might tune to scale back the quantity of false positives. This downside is solved by dynamic the classification rule to the subsequent l/k-rule: If l or additional messages among the k nearest neighbors of x area unit spam, classify x as spam, otherwise classify it as legitimate mail.

Our main objectives in this study are as follows:

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

IV. CONCLUSION

In Spam mail classification is major space of concern currently because it helps within the detection of unwanted emails and threats. thus, currently a day's most of the researcher's square measure operating during this space to search out the most effective classifier for detective work the spam mails. thus, a filter is needed with high accuracy to filter the unwanted mails or spam mails. during this paper we tend to focus on finding the most effective classifier for spam mail classification exploitation data processing techniques. thus, we tend to apply numerous classification algorithms on the given computer file set and check the results. From this study we tend to analyze that classifier works well once we engraft feature choice approach within the classification method that's the accuracy improved drastically once classifiers square measure applied on the reduced information set rather than the whole information set. As in projected the spam classification done on all parameters like scientific discipline, Previous history and content of shared uniform resource locator and information so the projected mechanism can help a great deal to travel improved spam mail detection.

V. REFERENCES

1. E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi and O. E. Ajibuwa, "Machine learning for email spam filtering: Review approaches and open analysis problems", Heliyon, vol. 5, no. 6, Jun. 2019.
2. W. Awad and S. ELseuofi, "Machine learning ways for spam E-Mail classification", Int. J. Computer. Sci. Inf. Technol., vol. 3, no. 1, pp. 173-184, Feb. 2011.
3. S. Mohammed, O. Mohammad, and J. Fiaidhi, "Classifying unsought bulk email (UBE) victimization Python machine learning techniques", Int. J. Hybrid Inf. Technol., vol. 6, no. 1,p.4355,2013https://www.researchgate.net/publication/236970412_Classifying_Unsolicited_Bulk_Email_UBE_using_Python_Machine_Learning Techniques.
4. A. Wijaya and A. Bisri, "Hybrid call tree and supplying regression classifier for email spam detection", Proc. 8th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE), pp. 1-4, Oct. 2016.
5. A. Géron, active Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. Newton, MA, USA: O'Reilly Media, 2019, Ch. 1.
6. (2019). 1. supervised Learning—Scikit-Learn zero.22.2 Documentation. Accessed: Gregorian calendar month. 9, 2019. [Online]. Available: https://scikit-learn.org/stable/supervised_learning.html
7. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, ''Scikit-learn: Machine learning in Python,'' J. Mach. Learn. Res., vol.12, pp.2825–2830, Oct.2011. [Online]. Available: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
8. S. Zhu and F. Chollet. (2019). in operation With RNNs. Accessed: Gregorian calendar month. 2, 2019.[Online].Available:https://keras.io/guides/working_with_rnns/
9. (2019). TensorFlow Core | Machine Learning for Beginners and consultants. Accessed: Nov. 2, 2019. [Online]. Available: https://www.tensorflow.org/overview
10. (2019). Spyder: The Scientific Python Development Environment— Documentation—Spyder three Documentation. Accessed: Nov. 2, 2019. [Online]. Available: https://docs.spyder-ide.org/
11. K. Agarwal and T. Kumar, "Email spam detection utilizing integrated approach of Naïve man of science, Proc. 2nd Int. Conf. Intel. Computer. management Syst. (ICICCS), pp. 685-690, Jun. 2018.
12. R. Belkebir and A. Guessoum, "A hybrid BSO-Chi2-SVM take aside to Arabic text indexing", Proc. ACS Int. Conf. Computer. Syst. Appl. (AICCSA), pp. 1-7, May 2013.
13. A. I. Taloba and S. S. I. Ismail, "An intelligent hybrid technique of genetic rule for E-Mail spam detection", Proc. 9th Int. Conf. Intel. Computer. Inf. Syst. (ICICIS), pp. 99-104, Dec. 2019.

14. 4. R. Karthika and P. Visalakshi, "A hybrid ACO primarily based all feature various manner for email spam classification", WSEAS Trans. Computer, vol. 14, pp. 171-177, 2015.

15. 5. S. L. Marie-Sainte and N. Alalyani, "Firefly rule primarily based all feature various for Arabic text classification", J. King Saud Univ.-Computer. Inf. Sci., vol. 32, no. 3, pp. 320-328, Mar. 2020.

16. 6. E. A. Natarajan, S. Subramanian, and K. Premalatha, "An enlarged cuckoo look for augmentation of bloom filter in spam filtering", Global J. Computer. Sci. Technol., vol. 12, no. 1, pp. 75-81, Jan. 2012.