MACHINE LEARNING BASED EMAIL SPAM FILTERING APPROACHES

Mr.MADHU SUDHANA RAO, MCA, DCA, DVR & Dr.Hima Shekar MIC College of Technology, A.P., India.

S.LAVANYA, Assistant Professor, Dept.of AI & IT, DVR & Dr.Hima Shekar MIC college of Technology, A.P., India.

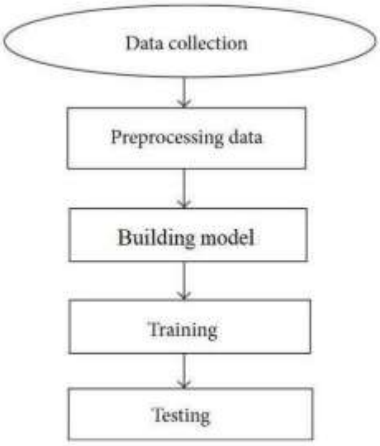
ABSTRACT:

In this study, This method normally analyses words, the occurrence, and distributions of words and phrases in the content of emails and used then use generated rules to filter
the incoming email spams Case Base Spam Filtering Method: Case base or sample base filtering is one
of the popular spam filtering methods. Firstly, all emails both non-spam and spam emails are extracted from each user's email using collection model. Subsequently, pre-processing steps are carried out to transform the email using client interface, feature extraction, and selection, grouping of email data, and evaluating the process. The data is then classified into two vector sets. Lastly, the machine learning algorithm is used to train datasets and test them to decide whether the incoming mails are spam or non-spam

INTRODUCTION

In recent times, unwanted commercial bulk emails called spam has become a huge problem on the internet. The person sending the spam messages is referred to as the spammer. Such a person gathers

email addresses from different websites, chatrooms, and viruses [1]. Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time [2]. The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the whole global email traffic [3]. Users who receive spam emails that they did not request find it very irritating. It is also resulted to untold financial loss to many users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails pretending to be from reputable companies with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number(BVN) and credit card numbers.



LITERATURE SURVEY

Sanz, Hidalgo, and Pérez [1] detailed the research issues related to email spams, in what way it affects users, and by what means users and providers can reduce it effects. The paper also enumerates the legal, economic, and technical measures used to mediate the email spams. They pointed out that based on technical measures, content analysis filters have been extensively used and proved to have reasonable percentage of accuracy and precision as a result, the review focused more on them, detailing how they work. The research work explained the organization and the procedure of many machine learning approaches utilized for the purpose of filtering email spams. However, the review did not cover recent research articles in this area as it was published in 2008 and comparative analysis of the different content filters was also missing.

Bhowmick and Hazarika [2] presented a broad review of some of the popular content-based e-mail spam filtering methods. The paper focused mostly on machine learning algorithms for spam filtering. They surveyed the important concepts, efforts, effectiveness, and the trend in spam filtering. They discussed the fundamentals of e-mail spam filtering, the changing nature of spam,the tricks of spammers to evade spam filters of e-mail service providers (ESPs), and also examined the popular machine learning techniques used in combating the menace of spam.

Laorden et al. [3] presented a detailed revision of the usefulness of anomaly discovery used for Email spam filtering that decreases the requirement of classifying email spam messages and only

works with the representation of single class of emails. The review contains a demonstration of the first anomaly based spam sieving method, an improvement of the method, which used a data minimization technique to the characterized dataset corpus to decrease processing phase while retaining recognition rates and an investigation of the appropriateness of selecting non-spam emails or spam as a demonstration of normality.

PROPOSEDSYSTEM

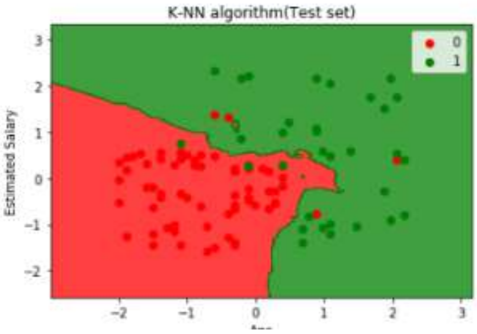
To effectively handle the threat posed by email spams, leading email providers such as Gmail, Yahoo mail and Outlook have employed the combination of different machine learning (ML) techniques such as Neural Networks in its spam filters. These ML techniques have the capacity to learn and identify spam mails and phishing messages by analyzing loads of such messages throughout a vast collection of computers. Since machine learning have the capacity to adapt to varying conditions, Gmail and Yahoo mail spam filters do more than just checking junk emails using pre-existing rules. They generate new rules themselves based on what they have learnt as they continue in their spam filtering operation. The machine learning model used by Google have now advanced to the point that it can detect and filter out spam and phishing emails with about 99.9 percent accuracy.

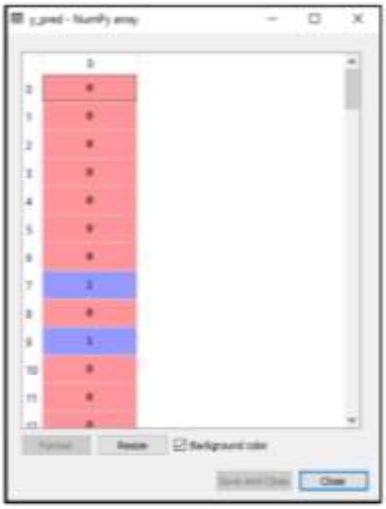
ADVANTAGES OF PROPOSED SYSTEM:

Though there are several email spam filtering methods in existence, the state-of-the-art approaches are discussed in this paper. We explained below the different categories of spam filtering techniques that have been widely applied to overcome the problem of email spam. Content Based Filtering Technique:

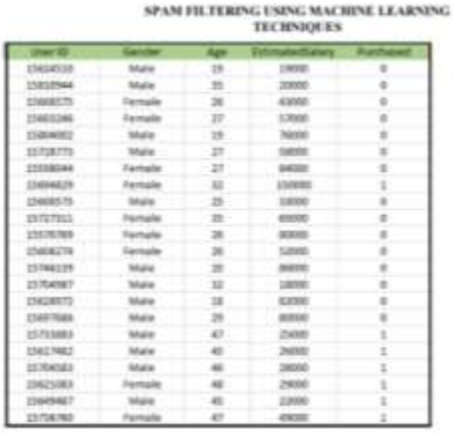
Content based filtering is usually used to create automatic filtering rules and to classify emails using machine learning approaches, such as Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbor, Neural Networks. This method normally analyses words, the occurrence, and distributions of words and phrases in the content of emails and used then use generated rules to filter the incoming email spams Case Base Spam Filtering Method: Case base or sample base filtering is one of the popular spam filtering methods. Firstly, all emails both non-spam and spam emails are extracted from each user's email using collection model.

SAMPLE RESULTS





	0
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	0
9	1
10	0
11	0
12	0



SPAM FILTERING USING MACHINE LEARNING TECHNIQUES

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1

CONCLUSION

In this project Our work mainlyfocused in the advancement of predictive models to achieve goodaccuracy in predicting valid flood prediction outcomes using supervised machine learning methods. The analysis of the results signify that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain. Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables

REFERENCES

https://www.sciencedirect.com/science/article/pii/S2405844018353404#:~:text=The%20machine%20learning%20model%20used,evading%20their%20email%20spam%20filter.

https://www.hindawi.com/journals/sp/2021/6508784/

https://towardsdatascience.com/email-spam-detection-1-2-b0e06a5c0472?gi=72abc66cf5e4

https://www.academia.edu/39186582/Loan_Default_Prediction_using_Machine_Learning_Techniques

[1]J.R.Douceur,“Thesybilattack,”inInternationalworkshoponpeerto-peersystems.Springer,2002,pp.251–260

[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, ``Detectingspammers on Twitter," in Proc. Collaboration, Electron. Messaging, Anti-Abuse Spam Conf. (CEAS), vol. 6, Jul. 2010, p. 12.

[3] S. Gharge, and M. Chavan, ``An integrated approach for malicious tweetsdetection using NLP," in Proc. Int. Conf. Inventive Commun. Comput.Technol. (ICICCT), Mar. 2017, pp. 435438.

[4] T. Wu, S. Wen, Y. Xiang, and W. Zhou, ``Twitter spam detection: Surveyof new approaches and comparative study," Comput. Secur., vol. 76,pp. 265284, Jul. 2018.

[5] S. J. Soman, ``A survey on behaviors exhibited by spammers in popularsocial media networks," in Proc. Int. Conf. Circuit, Power Comput. Tech- nol. (ICCPCT), Mar. 2016, pp. 16.

[6] A. Gupta, H. Lamba, and P. Kumaraguru, ``1.00 per RT #BostonMarathon# prayforboston: Analyzing fake content on Twitter," in Proc. eCrimeResearchers Summit (eCRS), 2013, pp. 112.

About authors:

Mrs. S.LAVANYA completed her Bachelor of Technology in Computer Science and Engineering. She completed her Masters of Technology in Computer Science and Engineering from JNTU KAKINADA UNIVERSITY. Currently working as an Assistant Professor in the department of CSE at DVR &amp; DR HS MIC COLLEGE OF TECHNOLOGY (Autonomous), Kanchikacherla (NTR Dist, AP). Her areas of interest are Data Mining, Cloud Computing and Machine Learning &amp; Networks.

Mr.MADHU SUDHANA RAO. T is an MCA student in the Department of DCA at DVR &amp; DR. HS MIC College of Technology (Autonomous), Kanchikacherla, NTR (DT), A.P. he completed B.Sc from Sir viswabarthi degree college , jaggayypet(NTR). His areas of interest include Machine Learning, Data Mining, Cloud Computing.