International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | June 2022 www.irjet.net p-ISSN: 2395-0072

E-Mail Spam Detection Using Supportive Vector Machine

Muskan Dhirwani[1], Kartikey Tiwari[2], Mandeep Singh Narula[3]

[1,2] *Students, Dept. of ECE, Jaypee Institute of Information Technology, Noida, India*
[3] *Assistant Professor, Dept. of ECE, Jaypee Institute of Information Technology, Noida, India*

Abstract -

With the rapid pace of growth of the internet it has been the most tedious job to classify the messages into ham and spam. Spam is defined as an unwanted message sent to somebody via email and messages. Many messages or emails often contain viruses, phishing material in order to break the privacy of the person by looking into its various confidential information. Many viruses try to get into the computer by this process of spam messages, so we have tried to build our project using the SVM process so that we could increase the spam detection feature of the model using neural networks in order to reduce the error. The features involved include regression, naive bayes and several other regression models.

Keywords: *CNN, unwanted messages, viruses, filtration.*

1. INTRODUCTION

Messages in this present phase has changed its recognition, gone are the days when people used to message their friends with old simple monotonous simple text. But today after so much advancement with the time email has come into the playground. After the advancement of the internet tools like Artificial Intelligence have more power to play and influence the masses. This power either could be positive or it could be negative. In recent times email has been used for advertisement click bait videos and for several other promotional activities. In this project we have tried to build a model using Machine Learning which could segregate the spam mails which are sent to us. An email consists of three parts. The 1st consists of the source address, beneficiary location and topic.

The 2nd comprises the main part or the body. The 3rd part might contain reports, pictures, sound or video documents. With the acceleration of the web and huge use of email there's an ascent inside the pace of spam messages. It consists of undesirable, excluded or useless messages. Programmers and other illegal persons try to break the sovereignty of the country using various methods to get into the inner details of the person like giving them clickbait messages, videos lure them with the spam messages. But recently engineers have developed methods and procedures to find these types of emails. The advantage of these types of secure models are many like they try to avoid the spam and clickbait videos which have been sent to the person in order to carry out the illegal activities like phishing and spying into the system. This type of model allows the model to make the stock of the harmful words which allows the user to know about the lethalness of the virus and other spams. The offensive emails are sent to the people with their standard activity on the internet. By researching these principles one can work on the basic working of the model so that and so forth such that they can work on the basics of the spam. The model tries to reduce the spam finding value by several percent. In the learning-based work we have tried to implement the model using neural networks which would try to increase the spam detection ability of the model by various percentages. These models have tried to incorporate spam subjectivity, language identification and several other keywords which would try to increase the efficiency of the model. Thus, in this paper we have tried to build a model which could allow us to do spam detection using models like linear regression, naïve bayes, SVM.

2. BACKGROUND

The paper revolves around the utility of the Support Vector Machine tool in Email detection. These days emails are often used to send important and crucial messages and information. But with the pace and growth of the internet comes the danger of cybercrime which involves techniques like phishing and bugs etc. These techniques are employed by the technique of spam mails. Just by sending spam mails and making users tempted to attempt it, this creates the problem for the user as it steals the important data, images, records and contacts from the user to the other person. This paper plans to use the analytics and model of SVM to carry out spam detection.

With the rapid growth and development of the internet and the importance of mobility, emails have now become the tool of daily use. In the earlier days emails were mostly messages of plain text but as of now it comes with the hyperlinks, videos, sound and the other tempting tricks. An email message is broadly classified into three parts. The first part consists of the source address, sender address and subject. The second includes the message and the information which contains the most important part of an email. The third could contain reports, pictures, sound or video archives. With the speed

increase of the web and enormous use there has been significant growth of spam messages in the name of email.

Software engineers and other illegal persons try to break the security of the network of the innocent person by the very use of spam. This very thing can be used to find out the important personal information of the person which can be used in order to cheat him/her. The use of the SVM algorithm model has been employed after using various others which shows the high accuracy rate.

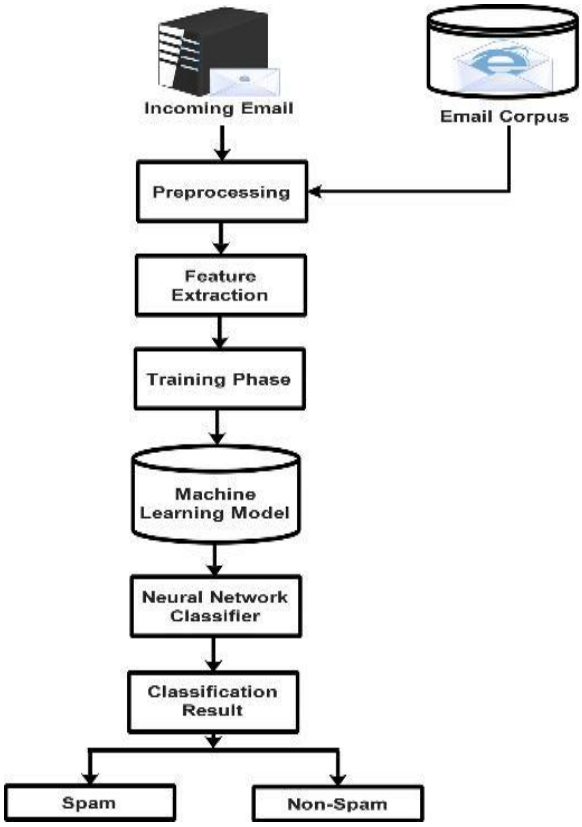

Fig 1. The architecture of the neural network (NN) Classifier

3. METHODOLOGY

We have tried to build our model using the SVM methodology. SVM has been declassified into the two models using the hyperplane concept which implies that data can be sorted into the two parts i.e. either in the linear vector format or in the nonlinear format. After which the hyperplane tries to divide it into two parts, the part closer to the hyperplane is the required thing. The model is of utmost efficiency if distance b/w two hyperplanes is maximum.

SVM has been involved in the model to find the appropriate result. The progress of SVM is essential because it involves the concept of the hyperplane and the mathematical idea which tries to build two lines both parallel to the hyperplane. Model would be a better model if it tries to maximise the distance between the two lines. In the model of artificial neural networks, it has been deployed so that it builds the linear model to accommodate the desired changes. The SVM has the ability to carry out double data access..

A. Pre-processing

It involves the process of lemmatization, somatization and tokenization.

-Evacuation of Nos

-Evacuation of a particular Symbol

-Evacuation of URLs

-Clearing HTML

-Word Stemming

B. Feature Extraction

Include Extraction is used to disparage the important and relevant highlights from the email body. The element changes the email into a 2 D vector space having a highlighted number. These elements are planned from the machine learning language list.

C. SVM Training

The email data set is classified by the combination of linear SVM and the non-linear SVM where linear SVM uses the hyperplane and non-linear SVM uses the kernel trick.

D. Test Classifier

The classifier has experimented with various preparation information for testing the accuracy of the classifier. The proposed arrangement succeeded with up to 98 % precision in classifying messages.

E. Test Email

After preparing a test model, a set classifier is applied to it . The classifier produces 0 and 1 , where 0 represents ham and 1 represents ham.

4. FINDINGS

Table -1: Accuracy in the system with different models

Model(Name)	Accuracy	Precision	F1 Score
Support Vector Machine Learning	0.98	0.99	0.98
Gaussian Naïve Byes	0.87	0.51	0.64
Multinomial Naïve Byes	0.96	0.96	0.96
Multinomial Naïve Byes	0.97	0.98	0.97
Logistic Regression	0.96	0.97	0.97

5. CONCLUSION

We have formed the SVM model and applied it to the dataset which shows the required result of SVM classification precision accuracy and the F1 score. After performing the model on the dataset it is showing that it has found out the spam and non-spams mails with accuracy of 98%. The model has also shown that it classifies the model into spam and non-spam by typing a particular message. It has also removed the abusive/derogatory/inappropriate emails from the inbox. This shows that the SVM model has utmost accuracy in the email spam detection as compared to other models.

The task of spam detection has been performed by using the appropriate model with the suitable algorithms and a model is now ready for further spam detection.

6. REFERENCES

[1] Sunil B. Rathod, Tareek M. Pattewar, "A Comparative Performance Evaluation of Content-Based Spam and Malicious URL Detection in E-mail", IEEE CGVIS 2015, pp: 49-54.

[2] Weimar Feng, Jianguo Sun, Liguo Zhang, Curling Cao, Qing Yang, "A Support Vector Machine based Naive Bayes Algorithm for Spam Filtering", IEEE 2016.

[3] Savita Teli, Santosh Kumar Biradar, "Effective Spam Detection Method for Email", IOSR Journal of Computer Science, pp: 68-75.

[4] Rohit Giyanani, Mukti Desai, "Spam Detection using Natural Language Processing", IOSR Journal of Computer Engineering, ISSN: 2278-0661, Volume 16, Issue 5, Sept-Oct 2014.

[5] Priyanka Sao, Kare Prashanthi, "Email Spam Classification Using Naive Bayesian Classifier", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4, Issue 6, June 2015.

[6] Omar Saad, Ashraf Darwish, and Ramadan Faraj, "A Survey of Machine Learning Techniques for Spam Filtering", IJCSNS International Journal of Computer Science and Network Security, Volume 12, February 2012.

[7] Akash Iyengar, G. Kalpana, Kalyankumar S., S. GunaNanshini, "Integrated Spam Detection for Multilingual Emails", International Conference of Information, Communication & Embedded System, IEEE 2017.

[8] S. Roy, A. Patra, S. Sau, K. Mandal, S. Kunar, "An Efficient Spam Filtering Techniques for Email Account", American Journal of Engineering Research (AJER), ISSN: 2320-0847, Volume 02, Issue 10, pp: 63-73, 2013.

[9] Rekha, Sandeep Negi, "A Review on Different Spam Detection Approaches", International Journal of Engineering Trends and Technology, Volume 11, May 2014.

[10]Nurul F. R, Norfaradilla W., Shahreen K., Hanayanti H, "Analysis of Naive Bayes Algorithm for Email spam Filtering across Multiple Datasets",International Research and Innovation Summit, 2017.

[11] W. A. Awad and S. M. Elseuofi, "Machine Learning Methods for Spam E-mail Classification", International Journal of Computer Science and Information Technology, Volume 3, February 2011.

[12] Kavitha, M., Manideep, Y., Vamsi Krishna, M., & Prabhuram, P. (2018). Speech controlled home mechanization framework using android gadgets. International Journal of Engineering and Technology(UAE), 7(1.1), 655-659.

[13] Modepalli Kavitha, Singaraju Srinivasulu, Kancharla Savitri, P. Sameera Afroze, P. Akhil Venkata Sai, S. Asrith l. (2019). "Garbage bin monitoring and management system using GSM." International Journal of Innovative and Exploring Engineering 8(7), pp. 2632- 2636.

[14] M. Kavitha, K Anvesh, P Arun Kumar, P Sravani .(2019). "IoT based home intrusion detection system." International Journal of Recent Technology and Engineering 7(6), pp. 694-698.

[15] Kavitha, M., et al. "Wireless Sensor Enabled Breast Self- Examination Assistance to Detect Abnormality." 2018 International Conference on Computer, Information and Telecommunication Systems (CITS). IEEE, 2018.

[16] Kavitha, Modepalli, P. Venkata Krishna, and V. Saritha. "Role of Imaging Modality in Premature Detection of Bosom Irregularity." Internet of Things and Personalized Healthcare Systems. Springer, Singapore, 2019. 81-92.

[17] A hybrid approach for securing the IoT devices published in IJITEE.Sai Prasanthi, Volume 8, Issue 4, 2019, Pages 147-151, ISSN: 22783075, CSE, Koneru Lakshmaiah Education Foundation, Guntur.

[18] I. Stuart, S.H. Cha, C. Tappert A neural network classifier for junk e-mail Document Analysis Systems VI, Springer Berlin Heidelberg (2004), pp. 442-450.