International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

EMAIL SPAM DETECTION USING MACHINE LEARNING AND PYTHON

Darshana Chaudhari[1], Deveshri Kolambe[2], Rajashri Patil [3], Sachin Puranik[4]

[1,2,3,4]U.G. Student, Department of Information Technology, SSBT's College of Engineering and Technology, Bambhori, Jalgaon, India

ABSTRACT

Nowadays, all the people are communicating official information through emails. Spam mails are the major issue on the internet. It is easy to send an email which contains spam message by the spammers. Spam fills our inbox with several irrelevant emails. Spammers can steal our sensitive information from our device like files, contact. Even we have the latest technology, it is challenging to detect spam emails. This paper aims to propose a Term Frequency Inverse Document Frequency (TFIDF) approach by implementing the Support Vector Machine algorithm. The results are compared in terms of the confusion matrix, accuracy, and precision. This approach gives an accuracy of 99.9% on training data and 98.2% on testing data achieved by using the Term Frequency Inverse Document Frequency (TFIDF) based Support Vector Machine(SVM) system.

Keywords: *Machine learning, phishing attack, spear phishing, spam detection, spam email, spam filtering, and Support vector machines, Naive Bayes.*

1. INTRODUCTION

The Internet has become a common thing in our lives. The same message sends multiple times which affects the organization financially and also irritates the receiving user. In this project, a Spam Mail Detection system is proposed will classify the given email as spam or ham email. Spam filtering mainly focuses on the content of the message. The classification algorithm classifies the given email based on the content. 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 harm. In recent times, unwanted industrial bulk emails known as spam has become an enormous drawback on the net. The person causing the spam messages is noted because the sender. Such an individual gathers email addresses from completely different websites, chatrooms, and viruses. Spam prevents the user from creating full and sensible use of your time, storage capability and network information measure. the massive volume of spam mails flowing through the pc networks have damaging effects on the memory house of email servers, communication information measure, central processing unit power and user time. The menace of spam email is on the rise on yearly basis and is to blame for over seventy-seven of the entire international email traffic. Users United Nations agency receive spam emails that they failed to request realize it terribly irritating. it's conjointly resulted to much loss to several users United Nations agency have fallen victim of web scams and different dishonest practices of spammers United Nations agency send emails pretence to be from honorable firms with the intention to influence people to disclose sensitive personal info like passwords, Bank Verification variety (BVN) and MasterCard number.

2. PRICE PREDICTION SYSTEM

In the paper the emphasis is on machine learning technique to predict the Price of the Crop using the Naïve Bayes Algorithm. The price of the crop is determined by recognizing the patterns in our training dataset which is given as one of the inputs to the Algorithm. The inputs values for the parameters (Yield, Rainfall, Minimum Support Price, and

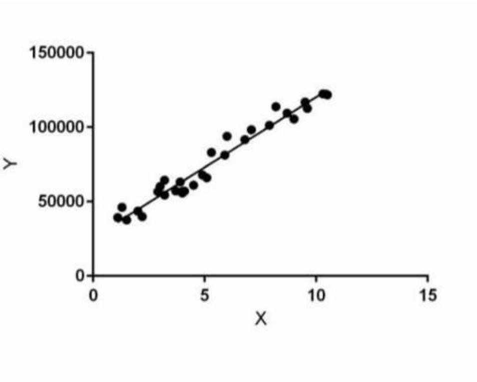


Fig 1 : Linear Regression

Fig 1 : Linear Regression

The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable.

3. SUPERVISED LEARNING

Naïve Baysian Algorithm Steps

Step1: Take the Raw data and extract the data which is required

Step 2: We need to get the probability of all the entities. [ex: sunny, windy, soil] Step 3: Now Applying the below formula

P = [n_c + (m*p)] / (n + m) Where:

- n_c = Count of the parameter when the price is same
- m = Number of parameters taken
- p = probability obtained
- n = Total number of prices
- P = price estimated Step 4: The result obtained will be assigned to a particular region and for a particular year Sample Example (Paddy) Parameters – P1, P2, P3 [m=3] Outcomes – 50rs, 80rs[p=1/2=0.5]

4. RELATED WORK

There are some research work that apply machine learning methods in e-mail classification, Muhammad N. Marsono, M. Watheq El-Kharashi, Fayez Gebali. They demonstrated that the naïve Bayes e-mail content classification could be adapted for layer-3 processing, without the need for reassembly. Suggestions on predetecting 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. They presented hardware architecture of na¨ıve Bayes inference engine for spam control using two class e-mail classification. That can classify more 117 millions 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 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 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). It integrates entities analogous to macrophages, B and T lymphocytes, modeling both the innate and the adaptive immune systems. An implementation of the algorithm was capable of identifying more than 99% of legitimate or SPAM messages in particular parameter configurations. It was compared to an optimized version of the naive Bayes classifier, which have been attained extremely high correct classification rates. It has been concluded 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.

5. METHODOLOGY

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 . This technique can be used 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, and make filtering decision based on the score. The statistic we are mostly interested for a token T is its spamminess (spam rating) , calculated as follows:

Most of the spam filtering techniques is based on text categorization methods. Thus filtering spam turns on a classification problem. In our work, rules are framed to extract feature vector from email. As the characteristics of discrimination are not well defined, it is more convenient to apply machine learning techniques. Three machine learning algorithms, C 4.5 Decision tree classifier, Multilayer There are number of rules framed by considering the various features that will aid to identify the spam messages effectively. Each rule performs a test on the email, and each rule has a score. When an email is processed, it is tested against each rule. For each rule found to be true for an email, the score associated with the rule is added to the overall score for that email. Once all the rules have been used, the total score for the email is compared to a threshold value. If the score exceeds the threshold, then the email is marked as spam and the others are classified as legitimate mail. In this work, the rules used are

$$S\,[T] = \frac{C_{Spam}(T)}{C_{Spam}(T)\ +\ C_{Ham}(T)}$$

Where CSpam(T) and CHam(T) are the number of spam or ham messages containing token T, respectively. To calculate the possibility for a message M with tokens {T1,......,TN}, one needs to combine the individual token's spamminess to evaluate the overall message spamminess. A simple way to make classifications is to calculate the product of individual token's spamminess and compare it with the product of individual token's hamminess

$$(\mathbf{H}\ [M] = \prod_{I=1}^{N}\ (\ 1\text{-}\ \mathbf{S}\ [T_{I}\ ]))$$

The message is considered spam if the overall spamminess product S[M] is larger than the hamminess product H[M]. The above description is used in the following algorithm [10]:

Stage 1. Training

```
        Parse each email into its constituent tokens
        Generate a probability for each token
                W $S[W] = Cspam(W) / (Cham(W) + Cspam(W))$
        store spamminess values to a database
```

Stage 2. Filtering

```
        For each message
         M while (M not end) do
         scan message for the next token Ti
         query the database for spamminess S(Ti)
        calculate accumulated message probabilities
         S[M] and H[M]
Calculate the overall message filtering indication by:
        $I[M] = f(S[M] , H[M])$
         f is a filter dependent function,
        such as  $I [M] = 1+S[M]-H[M]/ 2$
International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011 176
        if I[M] > threshold
                msg is marked as spam
        else
                msg is marked as non-spam
```

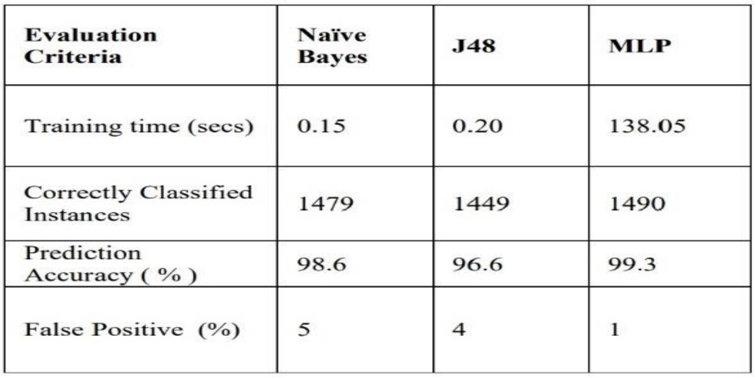
6. RESULT

The training dataset, spam and legitimate message corpus is generated from the mails that we received from our institute mail server for a period of six months. The mails are analyzed and 23 rules are identified that extremely ease the process of classifying the spam message. The corpus consists of 750 spam messages and 750 legitimate messages. From the corpus, the feature vectors are extracted by analyzing message header, keyword checking, white list/blacklist etc.

The class labels are designated as L and S to represent legitimate and spam message respectively. The machine learning techniques Naïve Bayes Classifier, C 4.5 Decision tree classifier, Multilayer Perceptron are used for training the dataset in WEKA environment. The training is carried out with the feature vectors extracted by analyzing each message header and keyword checking and whitelist/blacklist. The performance of the trained models is evaluated using 10-fold cross validation for its predictive accuracy. Predictive accuracy is used as a performance measure for email spam classification. The prediction accuracy is measured as the ratio of number of correctly classified instances in the test dataset and the total number of test cases. In spam filtering, false negatives just mean that some spam mails are classified as legitimate and moved to inbox. False positive mean that legitimate emails that get mistakenly identified as spam and moved to spam folder or discarded. For most users, missing legitimate email is an order of magnitude worse than receiving spam. The false positive rate of each classifier also considered to measure its performance. The performance of the classifiers are summarized in Table II and shown in Fig.2 and Fig.3.

TABLE II

COMPARATIVE RESULTS OF THE CLASSIFIERS



Evaluation Criteria	Naïve Bayes	J48	MLP
Training time (secs)	0.15	0.20	138.05
Correctly Classified Instances	1479	1449	1490
Prediction Accuracy (%)	98.6	96.6	99.3
False Positive (%)	5	4	1

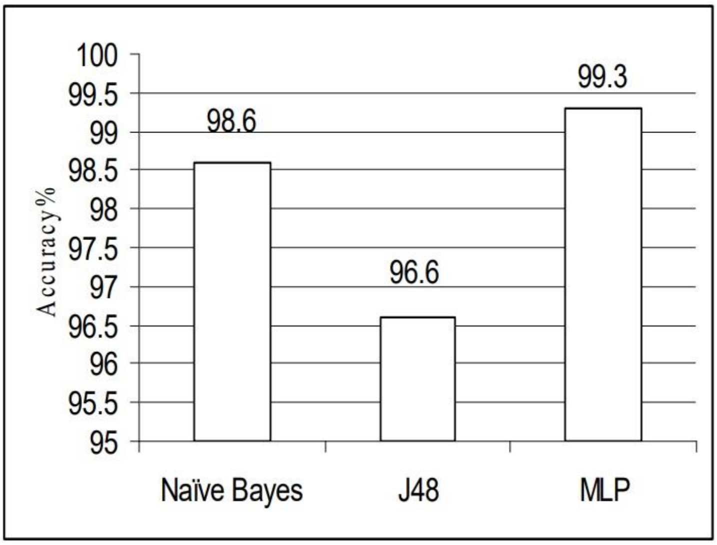


Fig 2: Classification Accuracy

The performance of the three models was evaluated based on the three criteria, the prediction accuracy, learning time and false positive rate. Multilayer perceptron predicts better than other algorithms.

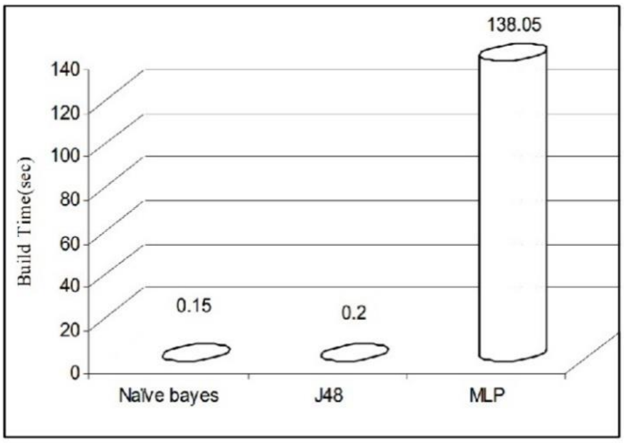


Fig 3: Learning Time of the Models

Multilayer perceptron, the neural network classifier consumes more time to build the model. The naivebayes, the probabilistic classifier and decision tree model tends to learn more rapidly for the given data set.

7. CONCLUSION

Email has been the most important medium of communication nowadays, through internet connectivity any message can be delivered to all aver the world. More than 270 billion emails are exchanged daily, about 57% of these are just spam emails. Spam emails, also known as non-self, are undesired commercial or malicious emails, which affects or hacks personal information like bank ,related to money or anything that causes destruction to single individual or a corporation or a group of people. Besides advertising, these may contain links to phishing or malware hosting websites set up to steal confidential information. Spam is a serious issue that is not just annoying to the end-users but also financially damaging and a security risk. Hence this system is designed in such a way that it detects unsolicited and unwanted emails and prevents them hence helping in reducing the spam message which would be of great benefit to individuals as well as to the company .In the future this system can be implemented by using different algorithms and also more features can be added to the existing system.

8. FUTURE WORK

Review spam detection is essential since it can ensure justice for the sellers and retain the trust of the buyer on the online stores. The algorithms developed so far have not been able to remove the requirement of manual checking of the reviews. Hence there is scope for complete automation of spam detection systems with maximum efficiency. With growing popularity of online stores, the competition also increases. The spammers get smarter day by day and spam reviews become untraceable. It is necessary to identify the spamming techniques in order to produce counter algorithms.

REFERENCES

[1] Suryawanshi Shubhangi, Goswami Anurag and Patil Pramod, Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers, pp. 69-74, 2019.

[2] A. Karim, S. Azam, B. Shanmugam, K. Krishnan and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection", IEEE Access, vol. 7, pp. 168261-168295.

[3] K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization", 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 685-690, 2018.

[4] Harisinghaney Anirudh, Aman Dixit Saurabh Gupta and Anuja Arora, "Text and image-based spam email classification using KNN Naïve Bayes and Reverse DBSCAN algorithm", Optimization Reliabilty and Information Technology (ICROIT), pp. 153-155, 2014.

[5] Mohamad Masurah and Ali Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification", Computer Communications and Control Technology (I4CT) 2015 International Conference .

[6] Shradhanjali Prof and Toran Verma, "E-Mail Spam Detection and Classification Using SVM and Feature Extraction", International Jouranl Of Advance Reasearch Ideas and Innovation In Technology, 2017, ISSN 2454-132X.

[7] W.A Awad and S.M ELseuofi, Machine Learning Methods for Spam E-Mail Classification.International Journal of Computer Science & Information Technology, 2011.

[8] A. K. Ameen and B. Kaya, "Spam detection in online social networks by deep learning", 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-4, 2018.

[9] D.D. Diren, S. Boran, I.H. Selvi and T. Hatipoglu, Root Cause Detection with an Ensemble Machine Learning Approach in the Multivariate Manufacturing Process, 2019.

[10] Tasnim Kabir, Abida Sanjana Shemonti and Atif Hasan Rahman, "Notice of Violation of IEEE Publication Principles: Species Identification Using Partial DNA Sequence: A Machine Learning Approach", 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering.