A detailed analysis on spam emails and detection using Machine Learning algorithms

Razia Sulthana A[1][0000−0001−5331−1310], Avani Verma[2][0000−0001−8077−4353], and Jaithunbi A K[3][0000−0002−2870−6609]

[1] Department of Computing and Mathematical Sciences, University of Greenwich, Old Royal Naval College, London, United Kingdom, SE10 9LS
razia.sulthana@greenwich.ac.uk
[2] Department of Computer Science, Birla Institute of Technology & Science, Pilani - Dubai Campus, United Arab Emirates- 345055
f20190077@dubai.bits-pilani.ac.in
[3] Department of Computer Science, RMD Engineering College, Kavaraipettai, TamilNadu, India, 601206
akj.cse@rmd.ac.in

Abstract. Spam Email is the unwanted junk and solicited email sent in bulk to the receivers, using botnets, spambots, or a network of infected computers. These spam emails can be phishing emails that trick users to get their sensitive information, download malware into the user devices or scam the users stealing confidential data. This paper shows a systematic analysis of spam and its types. It also details the procedure of how the spammers get the email addresses of the receivers. It analyses the problems with spamming. A detailed state of the art on spam filters and the factors that put an email into the spam or ham category is also explained. The paper also discusses spam filtering methods of Gmail, Yahoo, and Outlook. Finally, it brings out several solutions to detect spam using principles of Machine Learning and Data Mining.

Keywords: Spam email · Security breach · Naive bayes · Logistic Regression · Machine Learning

1 Introduction

Spam email is unwanted junk and unsolicited mail sent in bulk to the receiver through an email system like a network of infected computers and botnets. It can also be sent via text messages, phone calls and social media. It can be sent by businesses for commercial reasons. It can also be a malicious attempt to gain access to the user's computer. Since these emails are sent from botnets, they are very difficult to trace and stop.

The links or attachments in the mail might include malicious information. Generally, the hackers use the links or attachments to check the legitimacy of the email addresses or go to malicious websites or downloads which can install the malware in the computer. The users have their email addresses recognised by

spambots. Spambots are automated programs that search the internet for email addresses. Thus, spammers use spambots for generating an email distribution list. Emails are generally sent to millions of users. However, only a small number of users react to these emails.

The number of people communicating with each other online is increasing because of the internet. People depend on emails for general or business related issues. It is a very effective tool for communication as it saves cost and time. In recent years, emails are affected by attacks like spam emails, phishing emails etc. Spam floods receivers inboxes with mimicked messages, or with documents or links which can pass on malware to the device or can trick the receivers to reveal their sensitive information. Thus, spam filters are needed to avoid these. The spam filters should provide high accuracy and have minimal errors and should be efficient too. The objectives of the paper include:

• Understanding the meaning and types of spam emails.
• To analyse the working of spam email filters.
• To discuss case studies of Gmail, Yahoo and Outlook spam filters.
• To provide solutions to detect spam emails using Machine Learning (ML) algorithms (Naive bayes(NB) and Logistic Regression(LoR))..

2 Types of Spam

Spam can be used for the marketing of goods and services or can be malicious. Types of spam are:

1. Phishing Emails - these are sent by cyber attacker to many people which trick people into giving their personal information like bank and credit card details. Phishing is an online scam based on social engineering i.e., malicious activities performed through human interactions. The scammer creates links to click where users can put the sensitive information or download malware into the device.
2. Email Spoofing - the email spoofs an email that resembles the original so that the user can believe the authenticity. The message may be to request payment or to verify/ reset the account or update billing information..
3. Tech Support Scams - these emails explain that the user has some technical issue and provide the phone number of the tech support or a link to click. These emails mimic being a large and reputed company. If the user gives the details of the devices, they can be hacked.
4. Current Event Scams - these emails depend on the current news. For example, during Covid 19, scammers sent messages for work from home that paid in bitcoin or donations to fake organisations.
5. Advance Fee Scams - these scam emails promise a reward, generally financial if some amount of cash is provided in advance for some processing or transfer of money/goods. Once the user pays, they either ask more or disappear.

6. Malware Spam - it is a spam email that delivers malware to the devices. These emails have links or attachments. When the user clicks these, malware like Ransomware, Trojan Horses, Bots, viruses, Spyware etc are downloaded to the device. Generally, the attachments are in the form of a Word document, PowerPoint presentation or PDF file.

Spam filters can be implemented on all layers - firewalls in front of an email server or at Message Transfer Agent (MTA), email server to provide integrated anti-spam and antivirus solution which provides complete email protection at the network level. At the MDA level, spam filters can be installed.

2.1 Understanding how spammers gets address

Ways in which spammers get the email addresses:

1. There are thousands of companies that sell CDs containing millions of email addresses. These addresses can be easily formatted and copy-pasted in the 'To' section of the email. The companies get the email addresses from several primary sources like newsgroups and chat rooms. The users generally leave their email addresses in these groups. Software can be used to extract these screen names and email addresses.
2. Secondly, these email addresses can be found on the web. The '@' symbol can be searched on the internet to get the email addresses. This can be done by using a web crawler.
3. Thirdly, the spammer can create sites for winning the lottery in which the user has to type the email address. If they accept receiving the email newsletters, then their email addresses are sold to the spammer.
4. Next, performing a dictionary attack on the email hosting websites can also generate email addresses.

2.2 Problems with spamming

According to Statistica, the global daily spam volume from October 2020 to September 2021 was 1980.58 billion spam emails from 2346.05 billion emails, which accounted for 84.42%.The number of global daily spam volumes in July 2021 had the highest value of 283 billion spam emails out of 336.41 billion emails sent from October 2020 to September 2021. (Table 1 & Figure 1)

Every Internet Service Provider (ISP) pays to use the internet by the purchase of bandwidth. When the volume of spam that is directed to the ISP increases, the bandwidth becomes crowded, and this reduces the speed of internet access. To avoid this situation, the ISP pays to the filtering software or to increase the bandwidth. This expense is often passed to the buyers of ISP.

Some spam emails allow the user to remove themselves from the subscriber list but when the users respond to the email, they verify that their email accounts are active. This might lead to getting more spam emails.

Spam Emails	Total Emails
242.42	286.41
210.54	248.7
140.56	166.38
122.33	144.76
150.93	178.3
138.09	163.87
88.21	104.2
200.24	236.74
249.95	296.81
282.93	336.41
65.5	77.8
88.88	105.67

Table 1. Global daily spam volume from October 2020 to September 2021, Source: Statistica

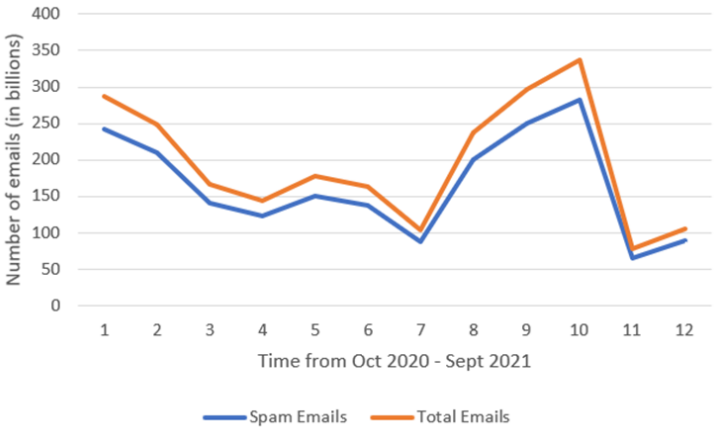

Fig. 1. Global daily spam volume from October 2020 to September 2021.Source: Statistica

2.3 Types of Spam Filters

1. Third-Party or Cloud-based and Gateway spam filters - several companies use cloud based or gateway spam filters for the suspicious inbound and outbound emails. These gateway spam filters are installed on servers. Whereas cloud-based spam filters are run on third-party servers. They are entirely digital. These cloud-based and gateway spam filters provide the network admin to have extra control over in and out traffic of the network.
2. Desktop spam filters - live on the user's device and allow for 1-1 configuration and personalisation. Ex. G-Lock, SpamCombat, Microsoft Smart Screen.
3. Email Service Provider (ESP) built-in spam filters - for Business to Customer or Business to Business senders, Google, Yahoo and Microsoft have inbuilt spam filters and inbox sorting technologies.

2.4 Factors to determine if email is spam or ham

A ham is generally an email that is not a spam.

1. Source Internet Protocol (IP) Address - If a specific IP address has received many complaints in the past, email from that address is more likely to be identified as spam. An email with a poor IP reputation might not be accepted by the server because the IP address is an important factor in delivering an email. Therefore, many companies set up dedicated IP addresses for their users.
   If there is no dedicated IP address, the emails will be sent through the marketing automation platform shared IP. Thus, the emails are sent from the same servers as the other customers through these platforms. If any user misbehaves, all other users who are using that platform are affected and the IP reputation decreases.
   For dedicated sending IP addresses, the reputation can be checked by using Sender Score, Talos Intelligence and Reputation Authority services. For shared IP, it is important to determine if the email is spam or ham, which will be decided by the domain reputation.
2. Sender's Domain - the ESP look for the sender's originating IP address, sending domain and the sender's alias. If emails of a company's domain are marked as spam, there is a high possibility that these emails are not a priority for the receivers. If ESP labels them as spam, these emails will be missed. The reputation of the domain is not good. Whereas, if the emails are whitelisted, the reputation of the domain is good.
3. Spam Traps - if emails are sent to the spam traps, the domain and IP reputation decreases. If an email account is not used for a specific period of time, the email providers disable it. The ESP might recycle it and convert it to a spam trap. The senders who send spam emails to that account will be fined. Therefore, such email addresses should be removed from the email list. ESP might put fake email addresses so that the bots find them and put them into the mailing list. If the spammers use such mailing lists, they'll be penalised and put on a blacklist.

4. Blacklists - are lists of IP addresses that are owned by known spammers or people who let spammers use their devices. Some of the known blacklists are Return Path Reputation Network Blacklist (RNBL), Sbl.spamhaus.org (SBL), SpamCop (SCBL) etc.
5. Sending Rate - emails can fail to reach the inbox because too many emails are being sent to that server at a time. The gateway filters allow the admin to rate control the bulk email deliveries. Also, if the email is sent to multiple contacts at the same domain, the email might not be delivered. Thus, spreading out of sending email over time can increase the deliverability. The emails can be delivered over a window over time with each recipient getting the email at the predicted time. Throttling follows send-time optimisation which reduces the probability of the email being labelled as bulk delivery.
6. Content - ESP sorts email using Content and IP. Recent spam filter models work on patterns rather than specific words to avoid. The content of the email plays a major role in user engagement. If the email has poorly qualified content, the user might directly mark it as a spam approach and ignore it. Thus, the email should be made keeping the text, images and HTML in mind.
7. Authentication - this is used for ESP to verify the sender's and to prevent spam from reaching the inbox. The emails that don't clear these protocols are considered spam. Types of authentication protocols :
   - Domain Keys Identified Mail (DKIM) - uses EDS to verify if the emails are from the actual domain or spoofed.
   - Sender Policy Framework (SPF) - lets the sender specify the authorisation of the mail servers to send email to the receiver's domain.
   - Domain-based Message Authentication, Reporting, and Conformance (DMARC) - gives options to receivers to handle emails if it fails the SPF and DKIM protocols. It also gives information about the senders from the domain.

2.5 Working of Gmail spam filter

More than 1 billion people use Gmail in a month. Gmail uses several rule-based filters, integrating tensor flow and artificial intelligence into the spam filters. It focuses on IP and Domain Reputation, User Engagement, Content and Sending History.

Gmail looks at both the domain and IP address of the senders to distinguish email between spam and ham. The algorithms check the user response when distinguishing between ham and spam. The content of the email - header, body, link, images etc determine if the email is spam or inbox. The filtering depends on the words that are blacklisted as spam words.

Gmail has a database of blacklisted domains. An email is first checked in this database. If the email or domain is not known, it checks if any links present in the email are malicious or not by comparing them with the database. It will also check for any spelling or grammatical errors by comparing the words in the email with the list of trigger words that are mostly featured in the spam emails.

2.6 Working of Outlook spam filter

Microsoft relies on Sender Reputation Data Network (SRDN) along with engagement, spam traps and complaints to filter spam. SRDN uses a panel of voters from different users to train the spam filters. Emails received can be resent asking the users to vote if the email sent was spam or ham. Higher spam votes will lead the future emails to mostly go to the spam folder. It is harder to lower the complaint rate by sending a large volume of emails using SRDN.

2.7 Working of Yahoo spam filter

Yahoo checks the IP address, Domain, Sender, and Uniform Resource Locator (URL) reputation, along with DKIM and DMARC protocols. If emails have a certain sending rate, and there is a sudden increase in activity, the email can be marked as spam. It follows the same practices as Gmail and Outlook.

3 Literature Review

Based on ML and Data Mining, the following literature review has been done (Table 2): In [16], P. Sharma et al. have focused on the ML [4, 13, 18] by implementing NB and J48 for spam email detection. The dataset is divided into different sets and given as input to each algorithm. Total three experiments are performed and the results obtained are compared in terms of Precision, Recall, Accuracy, F1 score, True Positive (TP) rate, True Negative (TN) rate, False Positive (FP) rate and False Negative (FN) rate. The two experiments are performed using individual Naive Bayes(NB) & J48 algorithms. In [8], P. Pandey et al have examined the ML strategies: NB, Support Vector Machines (SVM) relevance to the issue of spam email detection. Email filtration depends on the data classification approach. For data classification, choosing the best performing classifier is the base. Dataset used is of Ling Spam corpus. Firstly, data is accumulated and represented. Next, dimensionality is reduced by email feature choice.

In [15], M. Sethi et al. have proposed a work that focuses on Natural Language Processing (NLP) [3]. The technique used for detection is the NB and Artificial Neural Network ANN [19]. The steps involved are dataset reading and inspection, text preprocessing, feature set and vectorisation, pipeline. In [7], F. Martino et al. have given information about legitimacy to detect spam. The dataset is from Bruce Guenter Project. The algorithm used are NB, LoR, RFC, SVM. The methodology is first defining classes and features, building a dataset, and using vectors for feeding the classifier. In [9], L. Huang et al. have used NB for the classification of emails. The dataset is from Ling Spam Corpus. The methodology used is to preprocess data, searching common spam keywords. The advantage is that the testing has been carried out on spam encryptions.

In [2], N. Shah et al. have used LoR, k-Nearest Neighbour (K-NN) [14] and Decision Trees (DT) for spam detection. The dataset used is of SMS spam collection. There were 4900 ham samples and 672 spam samples. The advantage

is that the proposed method performance is good as compared with the existing state-of-the-art methods. The limitations is that the research is limited to few algorithms. The work can be improved by comparing more algorithms [23]. In [17], M. Singh et al. have used SVM [12, 1, 20] classifier for SMD. Non-linear SVM is used with two kernel functions - Linear and Gaussian kernel. The dataset is of Spam Assassin Public Corpus. In [5], O.E. Taylor et al. have used SVM and RFC models. The dataset used was of the UCL spam base. The methodology is used to preprocess data and split it into train and test data. It is then checked for accuracy by implementing the algorithms and finally, it is classified into ham/spam. The limitation is that only two algorithms were compared.

Reference	Algorithm	Performance metrics
[16]	Naive Bayes, J48	Precision, Accuracy, F1-score
[8]	Naive Bayes, Support Vector Machines, Logistic Regression	Accuracy
[15]	Naive Bayes, Artificial Neural Network	Precision, Accuracy, F1-score
[7]	Naive Bayes, Logistic Regression, Support Vector Machines	Accuracy
[2]	Naive Bayes	Accuracy

Table 2. Literature Analysis

In [6], A. Naem et al have used k-NN, SVM, Bagging, Boosting and approaches for spam email detection. Thke dataset used is CS-DMC2010 and SpamAssassin. The advantage is that the method gets a low number of selected features and archives a high degree of classification precision. Text mining is used to extract textual signatures in [10]. LoR and DT is applied for spam detection [22, 11, 21].

4 Analysis of Machine Learning algorithms

According to Literature Review, hybrid Naive Bayes and Logistic Regression model is most accurate. Naive Bayes is based on Bayes Theorem (Equation 1) with the assumption of strong independence. It is a Probability-Based Classifier. The combinational and frequency values of the dataset are calculated under a probability set. The class that is nearer to the rear end is picked by the classifier.

$$P(\frac{a}{b}) = \frac{P(\frac{b}{a})P(a)}{P(b)} \tag{1}$$

b = set of feature vectors
a = class variable
P(a/b) = posterior probability that depends on the likelihood of attribute value

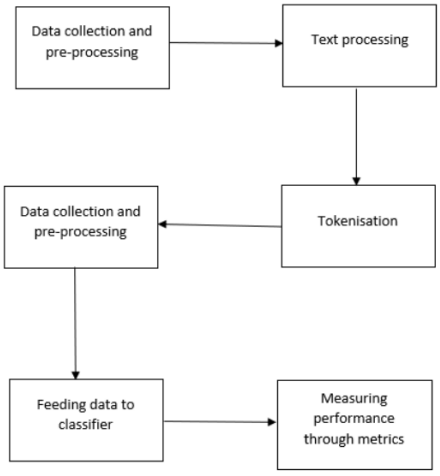

Fig. 2. Architecture Diagram

of class P(b/a)
P(a) = prior probability
P(b) = probability of known attribute value
Logistic is an analysis method to model the data and explain the relation between the Binary Response Variable and Explanatory Variable. The result is the probability of assigning a value to a particular class, which is in the range of 0 to 1.

1. Initially, data collection and pre-processing is done by removing undefined values, gaps and duplicates. This helps to reduce errors and improve the quality of classification.
2. Secondly, text processing is done to remove unwanted noise and characters like punctuation and numbers. This is done by converting all letters to lowercase, deleting numbers, and removing punctuation marks and stop words like prepositions and pronouns etc.
3. Tokenisation is performed by splitting sentences into words separated by a comma.
4. Finally, the quality of the classification is accessed. Model training is performed by the metrics namely Accuracy, Precision (Figure 2)

The dataset used for spam filtering is Ling spam dataset. It includes 1000 emails. The dataset is divided into 80:20 split ratio and are subjected to naive bayes and logistic regression. The training and testing results are shown in Table 3 and Table 4 respectively.

The results are shown in Figure 3. The precision, accuracy, F1 score of Naive bayes and Logistic Regression is recorded. The training and testing results shows a trivial difference which is because of the behavior of the model to the test data

Algorithm	Precision	Accuracy	F1-score
Naive Bayes	96.5	97.3	96.8
Logistic Regression	97.1	96.2	97.04

Table 3. Training Results for the Naive Bayes and Logistic Regression

Algorithm	Precision	Accuracy	F1-score
Naive Bayes	95.5	94.1	95.02
Logistic Regression	93.6	94.5	93.7

Table 4. Testing Results for the Naive Bayes and Logistic Regression

is unpredictable. However, the performance of the Naive bayes system is slightly better than logistic regressin in testing data and is vice versa in test data.

5 Conclusion

Spam email detection helps to detect the unwanted emails and threats. Many researchers are working in this field to find out the best classifier that is efficient and provides high accuracy in detecting spam emails and filtering them. Gmail, Outlook, Yahoo and other email service providers use ML algorithms integrated with AI and Data Mining to build their spam filter models. For ML based methods, NB and LoR prove to be the most efficient algorithms used to detect spam email. For Data Mining, the best method is RT proved to be the most accurate method with high accuracy.

References

1. Chauhan, A., Agarwal, A., Sulthana, R.: Genetic algorithm and ensemble learning aided text classification using support vector machines. International Journal of Advanced Computer Science and Applications **12**(8) (2021)

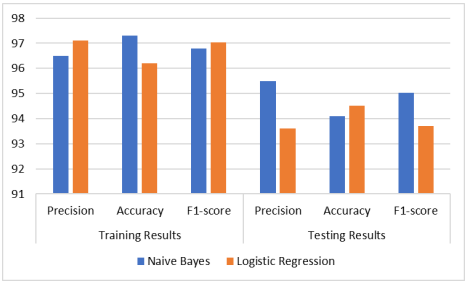

Fig. 3. Result analysis

2. GuangJun, L., Nazir, S., Khan, H.U., Haq, A.U.: Spam detection approach for secure mobile message communication using machine learning algorithms. Security and Communication Networks **2020** (2020)
3. Kontsewaya, Y., Antonov, E., Artamonov, A.: Evaluating the effectiveness of machine learning methods for spam detection. Procedia Computer Science **190**, 479–486 (2021)
4. Mathur, A., Sultana, R.: A study of machine learning algorithms in speech recognition and language identification system. In: Innovations in Computer Science and Engineering, pp. 503–513. Springer (2021)
5. Naem, A.A., Ghali, N.I., Saleh, A.A.: Antlion optimization and boosting classifier for spam email detection. Future Computing and Informatics Journal **3**(2), 436–442 (2018)
6. Naem, A.A., Ghali, N.I., Saleh, A.A.: Antlion optimization and boosting classifier for spam email detection. Future Computing and Informatics Journal **3**(2), 436–442 (2018)
7. Nidhya, M., Jayanthi, L., Sekar, R., Jeyabharathi, J., Poonam, M.: Analysis of machine learning algorithms for spam filtering. Annals of the Romanian Society for Cell Biology pp. 3469–3476 (2021)
8. Pandey, P., Agrawal, C., Ansari, T.N.: A hybrid algorithm for malicious spam detection in email through machine learning. Int J Appl Eng Res **13**(24), 16971–16979 (2018)
9. Peng, W., Huang, L., Jia, J., Ingram, E.: Enhancing the naive bayes spam filter through intelligent text modification detection. In: 2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE). pp. 849–854. IEEE (2018)
10. Qian, F., Pathak, A., Hu, Y.C., Mao, Z.M., Xie, Y.: A case for unsupervised-learning-based spam filtering. ACM SIGMETRICS performance evaluation review **38**(1), 367–368 (2010)
11. Rathi, M., Pareek, V.: Spam mail detection through data mining-a comparative performance analysis. International Journal of Modern Education & Computer Science **5**(12) (2013)
12. Razia, S.A., Pranav, R.: Predicting the import and export of commodities using support vector regression and long short-term prediction models. International Journal of Computing and Digital Systems **11**(1), 635–648 (2022)
13. Razia Sulthana, A., Mathur, A.: A state of art of machine learning algorithms applied over language identification and speech recognition models. In: International Virtual Conference on Industry 4.0. pp. 123–132. Springer (2021)
14. Sajedi, H., Parast, G.Z., Akbari, F.: Sms spam filtering using machine learning techniques: a survey. Machine Learning Research **1**(1), 1–14 (2016)
15. Sethi, M., Chandra, S., Chaudhary, V.: Email spam detection using machine learning and neural networks. Int. Res. J. Eng. Technol **8**, 349–355 (2021)
16. Sharma, P., Bhardwaj, U.: Machine learning based spam e-mail detection. International Journal of Intelligent Engineering and Systems **11**(3), 1–10 (2018)
17. Singh, M., Pamula, R., et al.: Email spam classification by support vector machine. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON). pp. 878–882. IEEE (2018)
18. Sulthana, A.R., Jaithunbi, A.: Varying combination of feature extraction and modified support vector machines based prediction of myocardial infarction. Evolving Systems pp. 1–18 (2022)

19. Sulthana, A.R., Jaithunbi, A., Ramesh, L.S.: Sentiment analysis in twitter data using data analytic techniques for predictive modelling. In: Journal of Physics: Conference Series. vol. 1000, p. 012130. IOP Publishing (2018)
20. Trivedi, S.K.: A study of machine learning classifiers for spam detection. In: 2016 4th international symposium on computational and business intelligence (ISCBI). pp. 176–180. IEEE (2016)
21. Vivekanandam, B., et al.: Spam email classification by hybrid feature selection with advanced machine learning algorithm–future perspective. Journal of Soft Computing Paradigm **4**(2), 58–68 (2022)
22. Wijaya, A., Bisri, A.: Hybrid decision tree and logistic regression classifier for email spam detection. In: 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE). pp. 1–4. IEEE (2016)
23. Yasin, A., Abuhasan, A.: An intelligent classification model for phishing email detection. arXiv preprint arXiv:1608.02196 (2016)