International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 01 | Jan 2021 www.irjet.net p-ISSN: 2395-0072

A LITERATURE SURVEY ON FILTERING EMAILS

Abdul Manaan Dar[1], Ankit Kumar[2], Abhinav Raj[3], Jay Prakash Kumar[4], Monika P[5]

[1]*Student, Dept. of Computer Science of Engineering, Dayananda Sagar College of Engineering, Karnataka, India*
[2]*Student, Dept. of Computer Science of Engineering, Dayananda Sagar College of Engineering, Karnataka, India*
[3]*Student, Dept. of Computer Science of Engineering, Dayananda Sagar College of Engineering, Karnataka, India*
[4]*Student, Dept. of Computer Science of Engineering, Dayananda Sagar College of Engineering, Karnataka, India*
[5]*Professor, Dept. of Computer Science of Engineering, Dayananda Sagar College of Engineering, Karnataka, India*

Abstract - *Nowadays, email messaging is a way with which people communicate important messages to each other using Internet. It's a very common way through which clients communicate among themselves formally. Now a days the extent to which these emails are sent has been increasing rapidly. Along with these emails, Spam emails are also sent in bulk through different platforms. These spam emails are usually difficult to recognize and it is the major problem that is being faced by the users. Spam consumes almost 98% of billions of emails sent every day. Due to the presence of different email filtering systems already present in the market, Spammers have become aware of these systems. Therefore, Spammers are trying different ways to send spam or junk mails to a number of users. One of them is by sending spam images and pdfs. For this kind of spam emails, presently there are not very effective systems present in the market. This paper illustrates a survey of different existing email classification system which can classify the email as ham or spam.*

***Key Words*: Spam, Email, Machine Learning, Naïve Bayes**

1. INTRODUCTION

Of all the different medium communication, email is extremely important medium now a days. It has been used widely for formal online communication. It can be accessed from any part of the world just with the help of internet connectivity. According to D Tschabitscher, number of active email accounts was 5 billion in 2017 and is increasing exponentially. He also stated that, everyday more than 270 billion Emails are exchanged, but the worst part of that is, out of that approximately 57 % emails are of no use as they are spam emails. Spam emails are creating a serious problem to the user as spammers flood the user's system with spam emails which results in storage problem, consumption of bandwidth and leads to decrease in performance of system.

Spam emails are called as junk emails or unsolicited message which is set by spammer through email. To make the email more secure and effective, appropriate email filtering is essential. Several types of researches have been done on email filtering, some acquired good accuracy but the progress is needed in this field. In order to avoid detection, spammers came with a new approach for sending spams to other users. It is included in the advertisements as the part of an embedded image file attachment in the form of.gif, .jpg, .png, etc. rather than body of the emails, hence by passing text-based spam filtering techniques. As we know that there are many techniques already there for email spam detection, our project aims for questing and analyzing the efficiency of the vital technique used for spam email detection from images and PDFs using Multinomial Naive Bayes' algorithm.

1.1 Machine Learning Models

[S.K Tuteja] (2016) [1]

The author has worked with different machine learning algorithms for email classification such as Neural Network (NN), Support Vector Machine (SVM), J48 Decision Tree based classifier, Naïve Bayes. The dataset used by the author was Spam Base dataset. In this paper work, the author didn't mention advantages and disadvantages of any algorithm.

[G. Mujtaba] [L. Shuib] [R. G. Raj] [N. Majeed] [M. A. Al- Garadi] (2017) [2]

Proposed the basic three steps which are common in every classification process. The first step is pre-processing in which the given text is converted into tokens and this step is also used for removal of stop words. The second step is learning process and, in this feature, set is built which is very much necessary for the classification of emails. The

last step is classification of email as ham or spam by using efficient algorithm. Algorithms like support vector machine, logistic regression, regression trees and random forest are considered for classification. They used the Phishing Corpus dataset and with the help of Bag of words as feature extraction approach classified the email as ham or spam. In his study, they did not mention the different tools for reduction methods for email classification.

[S. Ajaz] [M. T. Nafis] [V.Sharma] (2017) [3]

They collected email dataset from the online available websites and used Naïve Bayes for filtering of emails. He proposed a hybrid approach using secure hash method and Naive Bayes to filter email data but could not provide information regarding the misuse of storage resources and network bandwidth. By using Secure Hash Algorithm, the email is considered as a message M due to a generated function. The message M is further classified into S and L where L stands for ham email or genuine email and on the other hand S stands for spam email.

[Abdulhamid Muhammad Shafi] [M.S. Osho] [Ismaila] [J.K. Alhassan] (2018) [4]

They did performance analysis for different machine learning classification techniques such as Radial Basic Function (RBF) Network, Lazy Bayesian Rule, Random Tree, Bayesian Logistic Regression and J48. They did a comparison between all these given algorithms based on Precision, Recall, Root Mean Squared Error, F- Measure and Accuracy. They used the dataset from UCI Machine Learning Repository. For finding the precision and recall value, they applied the F-measure method. The highest F-measure was obtained by using Rotation forest algorithm and lowest for Naive Bayes algorithm. They used the Kappa Statistics for the statistical result and the best result was obtained for Rotation Forest Algorithm with 87.9. The best accuracy was obtained by using Rotation Forest algorithm with 94% accuracy and lowest accuracy was obtained by REP Tree algorithm with 89%. Other algorithms such as Naive Bayes gave the accuracy of 88.5% and J48 gave 92.3% accuracy.

[N.F. Rusland] [N.Wahid] [S.Kasim] [H.Hafit] (2017) [5]

Performed analysis on email classification on two different dataset by using Naïve Bayes algorithm based on the Accuracy, Precision, F-Measure and Recall. The process was divided into 3 steps. First step is data pre- processing in which all articles, conjunctions and undesired words is removed from the text. Next is the feature extraction followed by training of the Naive Bayes model. Based on the training of the model, it predicts whether the given text is ham or spam. By using Spam data Dataset, the author achieved an accuracy of 91.13% and for the other Spam Base dataset, accuracy achieved was 88%. By his analysis, the author concluded that the performance of Naïve Bayes algorithm is better on Spam data dataset compared to Spam Base.

[A.S Yuksel] [S.F. Cankaya] [I.S. Uncu] (2017) [6]

They compared Support Vector Machine (SVM) and Decision Tree for email filtering. The given dataset was divided into training set and testing set. Each of the model gets trained separately and based on its training, its accuracy is measured. The author made use of supervised learning for both the algorithms and obtained an accuracy of 92% on SVM and an accuracy of 82% on Decision Tree method. Based on his work, the author concluded that SVM performed better than Decision Tree.

[T. Verma] (2017) [7]

Proposed a method using SVM algorithm and feature extraction for filtering emails. This method consists of several steps such as Email Collection in which data is obtained from the dataset. After that it sent for pre-processing where unnecessary contents are removed and only desired content is sent for further process. Then the process of feature extraction followed by training of SVM model. The author used the dataset from Apache Public corpus. In the proposed solution, special symbols, HTML tags, URL and unnecessary alphabets were removed. The author used the Vocab file to map all the words from the dictionary. By using the SVM algorithm on pre- processed dataset, an accuracy of 98% was obtained.

[V.K Singh] [S. Bhardwaj] (2018) [8]

They worked on the solution for combining classification technique to get better result for spam filtering. The author took help of data mining and by using that collected all the information regarding success, current problems and previous failures of spam filtering. The method was based on binary classification where 1 was used for Spam email and 0 was used for Ham emails. They combined the 2 method that is Machine Learning and Knowledge Engineering for the filtering of emails. The performance of the proposed method was very poor on the combined KNN and SVM algorithm.

[Priti Sharma] [Uma Bhardwaj] (2017) [9]

Performed comparison between the Naive Bayes and J48 decision tree algorithm for classification of emails. They used the dataset of size 1000.They performed three experiments and based on the results, the algorithms are compared by evaluating the different performance parameters like accuracy, recall, precision, true negative rate, F-measure. First experiment performed by using Naive Bayes classifier and accuracy achieved was 83.5% along with precision value of 85.26% and recall value of 85.26%. Second experiment performed by using J48 decision tree classifier and accuracy achieved was 91.5% with precision value of 93.68% and recall value of 89%.

The third experiment was done by using hybrid bagged approach. In last experiment accuracy achieved was 87.5% with precision value of 89.47% and recall value of 85%. For future enhancement, the concept of boosting approach can be used as it might replace the weak classifier's learning features with the strong classifier's features.

[Manmohan Singh] [Rajendra Pamula] [Shudhanshu Kumar Shekhar] (2018) [10]

The author has compared the Gaussian Kernel and the Linear Kernel by using the Support Vector Machine algorithm for classification of emails. Linear separable problem is the one in which Linear Decision Boundary can be used to separate a class. For a particular problem, there can be several decision boundaries but a good and an efficient decision boundary is the one that perfectly fits the data given and also able to classify any new data. Gaussian boundary can be used where linear decision boundary is not effectively fitting to the given dataset. They used the spamTrain.mat dataset of size 4k for training purpose which contain both ham and spam emails. SpamTest.mat file containing 1k entry was used for testing purpose. Both the training and testing files are a subset of Spam Assassin Public Corpus. Along with the testing accuracy, they also focused on the training time and testing time in both approaches. Accuracy obtained by using Linear kernel was 98.5% with training time of 134 (in sec) and by using Gaussian Kernel, the accuracy was 97.1% with training time of 190(in sec).So based on the result, the author concluded that training time for linear kernel is far less than Gaussian kernel and also Linear Kernel has higher accuracy than that of Gaussian Kernel. Though the Gaussian kernel is more advanced and better fitting kernel than linear kernel but the dataset used has large number of features and thus a dataset having large number of features fit more better using linear kernel than Gaussian Kernel.

[Linda Huang] [Julia Jia] [Emma Ingram] [Wuxu Peng] (2018) [11]

Proposed a solution to increase the accuracy of Naive Bayes and reduce the false positive rate. Naive Bayes is a supervised machine learning algorithm which is based on Bayes theorem and can be used as a probabilistic model for classification of emails. Although Naive Bayes classifier provides higher accuracy but still spammers are able to bypass the filter by using leetspeak and diacritics. Leetspeak is a coded spelling system and language used in very informal communication on the internet, featuring letters combined with numbers or special characters in place of letters that they may resemble, and including inventive misspellings, jargon, and slang. Diacritic is a sign, such as an accent or cedilla, which when written above or below a letter indicates a difference in pronunciation from the same letter when unmarked or differently marked. They have done some modification in Naive Bayes to convert the symbols present in the text into possible letters and used a spell check to make sure that the corrected symbol is a word and then it is passed through the algorithm for classification. By doing this, they improved the accuracy from 23.9% to 62%.

[Prachi Gupta] [Ratnesh Kumar Dubey] [Dr. Sadhna Mishra] (2019) [12]

In this, they have compared the performance of Naive Bayes and Support Vector Machine algorithm for classification of emails. The dataset they have used consists of 5574 rows and 2 columns. One column is used for storing emails and other is used as label (Ham or Spam). Totally, they used 4 steps as Data Collection, Data Preprocessing, Data Transformation and Classification System for classification of emails. Data Pre-processing was used to clean the data and make it free from any kind of ambiguities, errors, redundancy. In Data Transformation, pre-processed data is converted into lowercase and converted into format as desired by the algorithm for classification. And at last desired attributes are identified and by using feature extraction, algorithm classifies the content into Ham or Spam. Accuracy obtained by using Naive Bayes was 99.49% and it was 86.35% by using Support Vector Machine. So, the author concluded that Naive Bayes algorithm performed exceptionally well as compared to SVM for classification of emails.

[U.K Sah] [N.Parmar] (2017) [13]

Proposed a method for classifying an email as Ham or Spam by using feature selection and also worked to improve the training time as well as the accuracy of the spam filtering model. They also performed a comparison between the Naive Bayes algorithm and Support Vector Machine. Based on the accuracy as well as the computation time of the algorithm for the given dataset. The whole process was divided into four steps. First was preparation of data in which the given dataset was divided into training set consisting of 702 mails and testing set with 260 mails. The second step was creation of word dictionary followed by the third step which was feature selection process by generating the feature vector matrix. The last step was to train the model and based on its training the model predicts the email as ham or spam. Based on the results obtained, the author concluded that Naive Bayes gives better accuracy in comparison with Support Vector Machine.

1.2 Pattern Matching Models

[D.Ruano-Ordas] [F.Fdez-Riverola] [J.R.Mendez] (2018) [14] They basically made use of regular expression to find a word or set of words showing some pattern. They did some modification in existing algorithm and developed an efficient algorithm named DiscoverRegex. This algorithm was dynamic in nature and was able to automatically produce regular expressions for a given dataset.

[A.S Aski] [N.K Sourati] (2016) [15] The author personally collected the spam or ham emails from various sources. They analyzed the collected dataset and carefully selected 23 features which were deciding factor for an email to be ham or spam according to them. They assigned some value for each of the criteria and based on the analysis, they fixed a threshold value. For each of the email for classification, total value was calculated and checked whether it is greater than or less than the threshold value and based on that, the result was given. That was not very effective as the study was done on a limited size dataset of just 750 emails.

Year	Reference Number	Evaluation Metrics	Dataset	Future Work
2016	[1]	Neural Network, Support Vector Machine, J48 Decision Tree	Spam Base Phishing Corpus	Algorithm can be used with the dataset having larger size
2017	[2]	Support Vector Machine, Logistic Regression, Regression Tree, Random Forest		
2018	[4]	Radial Basic function, Lazy Bayesian Rule, Random Tree, J48	UCI Machine Leaning Repository	Efficient Algorithm is required to achieve more accuracy
2017	[7]	Support Vector Machine	Apache Public Corpus	
2018	[8]	k-nearest neighbors, Support Vector Machine	Online Available Websites	Efficient method to achieve high accuracy, Acceptable Recall and Precision Value
2017	[9]	Naive Bayes, J48	Ling Spam Dataset	Concept of boosting approach can be used as it might replace the weak classifiers leaving features
2018	[10]	Gaussian Kernel, Linear Kernel using Support Vector Machine	Spam Assassin Public Corpus	
2018	[11]	Naive Bayes	Ling Spam Corpus	To Increase Speed and Efficiency and Also Detect other Forms of Email Messages
2016	[15]	Feature Extraction		Larger Size Dataset is Required

1.3 Spam Avoidance

We can deal SPAM in two ways, i.e. by checking and blocking spam from the originating place or we can use the other way which is to check and classify the mail as HAM or SPAM. Spammers target the servers which allow another server to use them as intermediate channel for forwarding messages. Botnets are the servers which are unattended servers with low security as mentioned. By changing continuously, the originating place and using botnets, it is very difficult for us to check for the mail for spam from the originating place itself. Many servers are black listed for spreading SPAM or it being used as for spreading SPAM. IP addresses of these black listed servers are black listed and distributed over mail servers. The mails which comes from these servers with which having blacklisted IP addresses is classified as SPAM without a second thought. As we don't trust the source it is without a second thought classified as SPAM. When spammers use open proxy servers it is very difficult to identify the source of the mail. Then comes the second method of checking mail and classifying it as HAM once email is already received at the mail server.

2. CONCLUSIONS

Spam emails have become a major concern for the internet community as it poses a threat to integrity and productivity of the users. Filtering of email is very much necessary for email communication. The accurate detection of spam emails is a big issue and many filtering methods have been proposed by various researchers.

After analyzing different papers given by different researchers, we observed as follows

- SVM algorithm was not able to give better result in terms of accuracy.
- Naïve Bayes has better performance than other algorithms such as Support Vector Machines (SVM's) and Decision Trees.
- Decision Tree Classifier was taking large memory space which is a great matter of concern for the researchers.
- Size of dataset used by many authors is very small and needs to be expanded.
- Most of the proposed model has four basic steps, preprocessing of data, feature extraction, training and testing.
- Some models used pattern matching technique also for classification by using regular expression.
- Spammers are now evolving and sending spam emails containing pictures and pdf to pass the filter.

REFERENCES

[1] S. K. Tuteja, "Classification Algorithms for Email Spam Filtering", 2016.
[2] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues", 2017.
[3] S. Ajaz, M. T. Nafis, and V. Sharma, "Spam Mail Detection Using Hybrid Secure Hash Based Naive Classifier, 2017.
[4] Shafi'i Muhammad Abdul Hamid, M. S., Osho, O., Ismaila, I., & Alhassan, J. K. "Comparative Analysis of Classification Algorithms for Email Spam Detection", 2018.
[5] Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H.. "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets", 2017.
[6] Yüksel, A. S., Cankaya, S. F., & Üncü, İ. S. "Design of a Machine Learning Based Predictive Analytics System for Spam Problem", 2017
Verma, T,"E-Mail Spam Detection and Classification Using SVM and Feature Extraction", 2017.
[7] Singh, V. K., & Bhardwaj, S,"Spam Mail Detection Using Classification Techniques and Global Training Set", 2018.
[8] Priti Sharma, Uma Bhardwaj,"Machine learning based Spam email detection",2017
[9] Manmohan Singh, Rajendra Pamula, Shudhanshu Kumar shekhar".Email Spam Classification by Support Vector Machine",2018.
[10] Linda Huang, Julia Jia,Emma Ingram, Wuxu Peng "Enhancing the Naive Bayes Spam Filtering through Intelligent Text Modification Detection",2018
[11] Prachi Gupta, Ratnesh Kumar Dubey,Dr. Sadhna Mishra,"Detecting Spam Emails/Sms Using Naive Bayes And Support Vector Machine",2019.
[12] Sah, U. K., & Parmar, N,"An approach for Malicious Spam Detection in Email with comparison of different classifiers", 2017.
[13] D. Ruano-Ordas, F. Fdez-Riverola, J.R Mendez,"Using evolutionary computation for discovering spam patterns from e-mail samples", 2018.
[14] A.S Aski, N.K Sourati,"Efficient algorithm to filter spam using machine learning techniques", 2016.