ISSN:0975-9646
KNS Prasad et al/ (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 12 (5) , 2021, 91-96

Email Spam Filtering using Machine Learning

Dr. KNS Prasad*, Rajani Surarapu, Lakshmi Mounika Pasumarthi, Bindu Bhargavi Sandaka, Harsha Uppalapati
Dept of computer science and engineering
Sasi Institute Of Technology & Engineering (Affiliated to JNTU Kakinada), Tadepalligudem, India
*ksnprasad@sasi.ac.in

***Abstract*— Internet has provided numerous modes for secure data transmission from one end station to a different, and email is one among those. The explanation behind its popular usage is its cost-effectiveness and facility for fast communication. Within the meantime, many undesirable emails are generated in a very bulk format for a monetary benefit called spam. Despite the very fact that individuals have the power to promptly recognize an email as spam, performing such task may waste time. Discussion on general email spam filtering process, and therefore the various efforts by different researchers in combating spam through the utilization machine learning techniques was done. Our review gives knowledge to the user about the fake e-mails and relevant e-mails and to classify that mail spam or not by using Naïve Bayes Algorithm.**

Keywords— Spam, Spam Filtering, Machine Learning, Naïve Bayes Algorithm.

I. INTRODUCTION

Email system is one in all the foremost powerful communication systems for transmitting the user's information from one to a different. It includes not only text but also images, files, etc. This platform helps the user in saving an enormous amount of your time and money compared with out dated techniques like telegrams, etc. Nowadays, approximately 281 billion emails are transmitted everywhere the globe in our day-to-day life. One of the threats to such quite platform is spam, where thousands of unwanted and unintended emails are transmitted daily. Spams are nothing but unwanted emails that are transmitted to various users having no prior permission. It should exist in various ways like unwanted and unwarranted advertising of products or services. In recent years, there has been tremendous growth within the volume of spam. In a very similar way, authorized, perfect and flawless emails are called ham.

Spam email comes in different ways. Many are just unwanted messages aiming to draw attention to a cause or spread the false information. Some of them are whaling emails with the intent of luring the recipient into clicking on a spiteful link or downloading a malware. The one thing they have in common is that they are unrelated to the needs of the recipient. A spam-detector algorithm must find a way to filter out spam while and at the same time it should avoid flagging genuine messages that users want to see in their inbox. And it must do it in a way that can match develop trends such as panic caused from pandemics, election news, sudden interest in crypto currencies and others.

Spam filters use "heuristics" methods, which suggests that every email message is subjected to thousands of predefined rules (algorithms). Each rule assigns a numerical score to the probability of the message being spam, and if the score passes a particular threshold the e-mail is flagged as spam and blocked from going further. There are different types of spam filters: Content filters – parse the content of messages, scanning for words that are commonly utilized in spam emails.

Header filters – examine the e-mail header source to look for suspicious information (such as spammer email addresses).

Blocklist filters – It can stop emails that come from a blocklist of suspicious IP addresses. Some filters can go further and check the IP reputation of the IP address.

Rules-based filters – can apply customized rules designed by the organization to exclude emails from specific senders, or emails containing specific words in their subject line or body.

Table 1 : Spam Categories

Categories	Descriptions
Health	The spam of fake medications
Promotional products	The spam of fake fashion items like clothes bags and watches
Adult content	The spam of adult content of pornography and prostitution
Finance & marketing	The spam of stock kiting, tax solutions, and loan packages
Phishing	The spam of phishing or fraud

II. RELATED WORK

There is a rapid increase within the interest being shown by the world research community on email spam filtering. During this section, we present similar reviews that are presented within the literature during this domain.

Lueg [10] presented a survey to explore the gaps in whether the data filtering and data retrieval technology may be applied to postulate Email spam detection in an exceedingly logical, theoretically grounded manner, so as to facilitate

the introduction of spam filtering technique that might be operational in an efficient way. However, the survey failed to present the small print of the Machine learning algorithms, the simulation tools, the publicly available datasets and also the architecture of the e-mail spam environment.

Wang [10] reviewed the various techniques wont to filter unsolicited spam emails. The paper also to classified email spams into various hierarchical folders, and instinctively regulate the tasks needed to give a response to an email message. However, a number of the restrictions of the criticism are that; machine learning techniques, email spam architecture, comparative analysis of previous algorithms and also the simulation environment were all not covered.

Cormack [11] previously stated that the spam filtering algorithms up to 2008 with specific significance on efficiency of the given systems. The most concentration of the review is to explore the relationships between email spam filtering with other spam filtering systems in communication and storage media. The paper also concentrate on the properties of email spams, including the wants user's information and therefore the task of the spam sieve as a constituent of an oversized and sophisticated data system. However, certain important components of spam filters weren't considered within the survey.

Bhowmick and Hazarika [13] presented a broad review of a number of the popular content-based e-mail spam filtering methods. The paper focused totally on machine learning algorithms for spam filtering. They surveyed the important concepts, efforts, effectiveness, and therefore the trend in spam filtering. They discussed the basics of e-mail spam filtering, the changing the nature of spam and the tricks of spammers to eradicate spam filters of e-mail service providers (ESPs), and also tested the popular machine learning techniques employed in combating the menace of spam.

Laorden [14] et al. presented an in depth revision of the usefulness of anomaly discovery used for Email spam filtering that decreases the need of classifying email spam messages and only works with the representation of single class of emails. The review contains an illustration of the primary anomaly based spam sieving method, an improvement of the tactic, which takes an information minimization technique to the characterized dataset corpus to reduce processing phase while retaining rates and an investigation of the appropriateness of choosing non-spam emails or spam as an illustration of normality.

III. PROPOSED WORK

Architecture:

Spam filtering is aimed toward reducing to the barest minimum the amount of unsolicited emails. Email filtering is that the processing of emails to rearrange it in accordance to some definite standards. Mail filters are generally accustomed manage incoming mails, filter spam emails, detect and eliminate mails that contain any malicious codes like virus, trojan or malware. The working of email is influenced by some basic protocols which include the SMTP. A number of the widely used Mail User Agents (MUAs) are Mutt, Elm, Eudora, Microsoft Outlook, Pine, Mozilla Thunderbird, IBM notes, K mail, and Balsa. They're email clients that assists the user to read and compose emails. Spam filters will be deployed at strategic places in both clients and servers.

Spam filters are deployed by many Internet Service Providers (ISPs) at every layer of the network, ahead of email server or at mail relay where there's the presence of firewall. The firewall is also a network security system that monitors and manages the incoming and outgoing network traffic supported on predetermined security rules. The e-mail server is an incorporated anti-spam and anti-virus solution providing comprehensive precautions for email at the network perimeter. Filters may be implemented in clients, where they'll mounted as add-ons i computers to function as intermediary between some endpoint devices.. Filters block unsolicited or suspicious emails that are a threat to the safety of network from attending to the pc system. Also, at the e-mail level, the user can have a customized spam filter that may block spam emails in accordance with some set conditions.

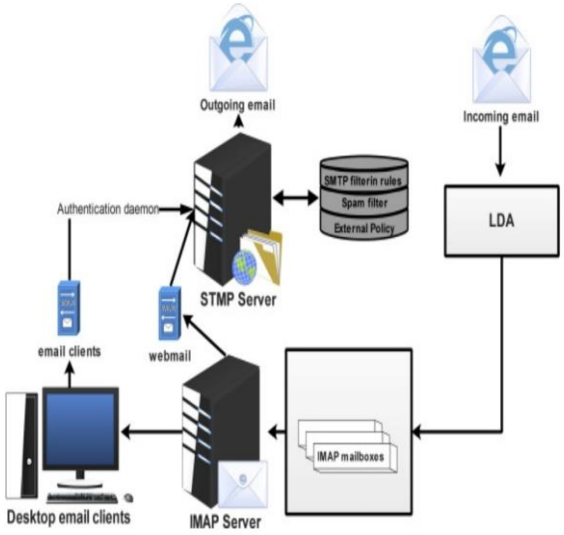

Fig 1: Architecture of Spam Filtering

METHODS:

Of recent, spam mail classification is generally handled by machine learning (ML) algorithms intended to differentiate between spam and non-spam messages. Machine learning algorithms can achieve this by using an automatic technique and adaptive technique. instead of betting on hand-coded rules that are prone to the perpetually varying characteristics of spam messages, ML methods have the capacity to get information from a group of messages provided, so use the acquired information to classify new messages that it just received. ML algorithms have the capacity to perform better supported their experience. During this section we are going to review a number of the foremost popular machine learning methods that are applied to spam detection.

Neural Network:

Artificial Neural Networks are groups of easy processing units which are interconnected, and communicate with each other by means of a large number of weighted connections.

Each of the units accepts input from the neighboring units and external sources and calculates the output that's transmitted to other neighbors. The medium for fine-tuning the weights of the connections is additionally made available. Neural networks are potent algorithm for solving any machine-learning problem that needs classification because of their resourcefulness, they're evolving as a serious tool within the machine-learning researcher's set of tools. Nevertheless, neural networks don't seem to be commonly utilized in the detection of spam email in concert may well envisage. As another, nearly all state-of-the-art spam filters use naïve Bayes classifiers. This is often due primarily to Paul Graham's well-known work titled "A Plan for Spam." Naïve Bayes is a wonderful method for spam classification with high accuracy (99.99+%) and an occasional false-positive rate.

More research efforts must be focused on the efficacy of the network across datasets rather than the appropriateness of diverse network designs for the work. There are generally three types of units.

•Input Unit: This unit receives signals from outside sources.
•Output Unit: This unit sends the data outside to the network.
•Hidden Unit: This unit both receives and sends signals within the network.

Algorithm : Perceptron Neural Network algorithm for Email Spam Classification

1: **Input** Sample email message dataset
2: Initialize w and b (to random values or to 0).
3: Find a training sample of messages (x,c) for which sign $(w^T x + b)$.
4: **if** there is no such sample, **then**
5: Training is completed
6: Store the final w and stop.
7: **else**
8: update (w,b): $w = w + cx$,
9: $b = b + c$
10: **go to** step 8
11: **end if**
12: Determine email message class as sign $(w^T x + b)$
13: **return** Final Email Message Classification (Spam/Non-spam email)
14:**end**

Support Vector Machines (SVM):

SVM are supervised learning algorithms that are proven to perform better than another attendant learning algorithms. SVM could be a group of algorithms proposed by for solving classification and regression problems. SVM finds application and provides solution to quadratic programming problems that have inequality constraints and linear equality by differentiating different groups by means of a hyper plane. It takes full advantage of the boundary. Though the SVM won't be as fast as other classification methods, the algorithm draws its strength from its high accuracy thanks to its capacity to model multidimensional borderlines that aren't sequential or straightforward. SVM isn't easily liable to a situation where a model is disproportionately complex like having numerous parameters comparative to the quantity of observations. These qualities make SVM the perfect algorithm for application within the areas of digital handwriting recognition, text categorization, speaker recognition, and so on. We explain briefly about the binary C-SVM classifier which was explained. Here C denote the price parameter to control modeling error which arises when a function is just too closely suited a limited set of knowledge points by penalizing the error.

Algorithm : Support Vector Machine (SVM) algorithm
1: **Input** Sample Email Message x to classify
2: A training set S, a kernel function, {c1, c2, …cnum} and {γ1, γ2, …γnum}.
3: Number of nearest neighbours k.
4: for i = 1 to num
5: set C=Ci;
6: for j = 1 to q
7: set γ=γ;
8: produce a trained SVM classifier f (x) through the current merger parameter (C, γ);
9: if (f (x) is the first produced discriminant function) then
10: keep f (x) as the most ideal SVM classifier f∗(x);
11: else
12: compare classifier f (x) and the current best SVM classifier f∗(x) using k-fold cross-validation
13: keep classifier with a better accuracy.
14: end if
15: end for
16: end for
17: return Final Email Message Classification (Spam/Non-spam email)
18: end

Naive Bayes algorithm:

Bayesian classification exemplifies a supervised learning technique and at the identical time a statistical technique for classification. It acts as a fundamental probabilistic model and allow us to seize ambiguity about the model in an ethical way by influencing the possibilities of the result. Bayesian classification is known as after Thomas Bayes, who proposed the algorithm. The classification offers practical learning algorithms and former knowledge and experimental data may be merged. It computes exact likelihoods for postulation and it's robust to noise in computer file. A Naive Bayes classifier may be a straightforward probabilistic classifier that's founded on Bayes theorem with sound assumptions that are independent in nature. The notion of sophistication restrictive autonomy was created to form computation easier, and is that the basis of tagging the algorithm 'naïve'. Nevertheless, the algorithm is effective and really robust. It performs a bit like other supervised learning algorithms.

Bayes Theorem: Prob(B given A) = Prob(A and B)/Prob(A)

Navie Bayes classifier:

During this research, Naıve Bayes classifier use bag of words features to spot spam e-mail and a text is representing as the bag of its word. The bag of words is often employed in methods of document classification, where the frequency of occurrence of every word is employed as a feature for

training classifier. This bag of words features are included within the chosen datasets. Naıve Bayes technique used Bayes theorem to work out that probabilities spam e-mail. Example, suppose that we all know exactly, that the word Free could never occur in a very non-spam e-mail. Then, after we saw a message containing this word, we could tell obviously that were spam email. Bayesian spam filters have learned a really high spam probability for the words like Free and Viagra, but an awfully low spam probability for words seen in non-spam e-mail, such as the names of friend and friend.

Algorithm : Naïve Bayes Classification Algorithm:

1: **Input** Email Message dataset
2: Parse each email into its component tokens
3: Compute probability for each token S [W] = $C_{spam}(W)/(C_{ham}(W) + C_{spam}(W))$
4: Store spamminess values to a database
5:**for** each message M **do**
6: **while** (M not end) **do**
7: scan message for the next token Ti
8: query the database for spamminess S(Ti)
9: compute probabilities of message collected S [M] and H [M]
10: compute the total message filtering signal by: I [M] = f (S [M], H [M])
11: $I[M] = \frac{I+S[M]-H[M]}{2}$
12: **if** I [M] > threshold **then**
13: msg is labeled as spam
14: **else**
15: msg is labeled as non-spam
16: **end if**
17: **end while**
18: **end for** 19: **return** Final Email Message Classification (Spam/Valid email)
20: **end**

Email Spam Filtering Process:
An email message is made up of two major components they are the header and the body. The header is the area where we have broad information about the content of the email. It has subject, sender and receiver. The body is the heart of the email. It may include information that does not have a pre-defined data. For example web page, audio, video, analog data, images, files, and HTML markup. The email header contains the fields such as sender's address, the recipient's address, timestamp which indicate when the message was sent by the intermediary servers to the Message Transport Agents (MTAs) that function as an office for organizing mails. The header line starts with "From" and it goes through some modification whenever it moves from one server to another through an in-between server. Headers allow the user to know about the route the email passes through, and the time taken by each server to go through the mail.
The necessary process that has to be observed within the mining of information from an email message were categorized in to following:

Adding corpus: It had been the first stage that's executed whenever an incoming mail is received. This step includes a dataset which undergoes through test and train processing.
Tokenization: this can be a process that removes the words within the body of an email. This is often the place where words are transformed in to meaningful message. It takes the e-mail and divides it into a sequence of representative symbols called tokens. Guzella and Caminhas are the methods of replacing information with distinctive identification symbols will make it free of all the characteristics and words from the e-mail exclusive of taking under consideration the meaning without losing their security.
Feature selection: Sequel to the pre-processing stage is that the feature selection phase. Feature selection a form of reduction within the measure of spatial coverage that effectively exemplifies fascinating fragments of email message as a compressed feature vector. The technique is helpful when the dimensions of the message is large and a condensed feature representation is required to form the task of text or image matching snappy. fee fraud, including inheritance, lottery, visa and customs-clearance scams, military scams, Ads for miscellaneous external sites, earning money through “work-from-home” jobs, online shopping, pleading and gift requests, business proposals et al. a number of the foremost important features for spam filtering include: Message body and subject, Volume of the message, Occurrence count of words, Circadian patterns of the message, Sex and country, Recipient age, Recipient replied (indicates whether the recipient replied to the message), Adult content and Bag of words from the message content. Sender Account Features used for spam filtering include: Sender Country (The distribution of states as stated by users on their profile and as revealed by their IP address), Sender Email, Sender IP address, Sender Reputation, Sender & Recipient Age. The low bothered features are: Geographical distance between sender and receiver, Sender's date of birth, Username and password of the sender, Account lifespan, Sex of sender and Age of recipient. The popularity of spam e-mails with minimum number of features is very important seeable of computational complexity and time. Feature selection involves processes like, noise removal, Stemming and stop word removal steps.

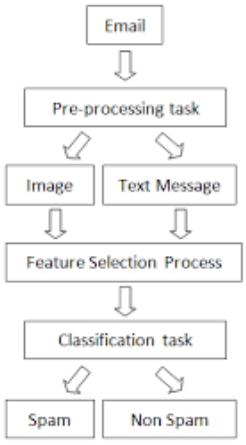

Fig 2: Process for email spam filtering

IV. RESULTS

In this paper, we explain the classification of unwanted emails to identify spam and not spam. For this we use the Naïve Bayesian Classifier. In this paper, we created a custom dataset to run this experiment. In our data set, we took some emails in which we classify the dataset into train set and test set in the ratio of 80% and 20%.

```
df = pd.read_csv('emails.csv')
df.head(5)
```

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny i...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1

Fig 3: Description of dataset

Evaluation of the model on training set:

In this section we find out the precision, recall, f1-score, support and accuracy for the training set. We also got the confusion matrix for training set which makes easy for the system to analyze. Here we got the accuracy with 99.7% with training set.

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      3457
           1       0.99      1.00      0.99      1099

    accuracy                           1.00      4556
   macro avg       0.99      1.00      1.00      4556
weighted avg       1.00      1.00      1.00      4556

Confusion Matrix:
 [[3445   12]
 [   1 1098]]

Accuracy:  0.9971466198419666
```

Fig 4: Evaluation of the model on training set

Evaluation of the model on testing set:

Now test the model on the test dataset (xtest & ytest) by printing the precision, recall, f1-score, support and accuracy for the training set.

```
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       870
           1       0.97      1.00      0.98       269

    accuracy                           0.99      1139
   macro avg       0.98      0.99      0.99      1139
weighted avg       0.99      0.99      0.99      1139

Confusion Matrix:
 [[862   8]
 [  1 268]]

Accuracy:  0.9920983318700615
```

Fig 5: Evaluation of the model on testing set

The classifier accurately identified the email messages as spam or not spam with 99.2% accuracy on the test data.

V. DISCUSSIONS

Majority of the present email spam filters doesn't possess the capacity to incrementally learn in real-time. Conventional spam email classification techniques aren't any longer viable to cope in real time environment that's characterized by evolving data streams and concept drift. Failure of the many spam filters to cut back their false positive rate. Development of more efficient image spam filters. Most spam filters can only classify the text in spam messages. However, many savvy spammers send spam email as text embedded in a picture (stego image) thereby making the spam email to evade detection from filters. By applying semantic web and ontology to spam email filtering there is need to develop scalable, adapted and integrated filters. Some filters lack the capacity to dynamically update the feature space. Majority of the prevailing spam filters are unable to incrementally add or delete features without re-creating the model totally to stay au courant current trends in email spam filtering. The need to use deep learning to spam filtering so as to use its numerous processing layers and plenty of levels of abstraction to find out representations of knowledge.

VI. CONCLUSION AND FUTURE WORKS

In the classification of spam, the most source of concern is that the classification of e-mail and unwanted threats. So today, most researchers are working during this area to seek out the simplest classifier to detect spam. Therefore, you wish a filter with great precision to strain spam mails or spam mails. Using this article, we've got focused on finding the most effective classifier for spam classification using data processing techniques. Therefore, we apply different classification algorithms within the given computer file set and verify the results. During this study, we analyzed that classifiers work well after we incorporate the feature selection approach into the classification process, International Journal of Scientific Development and Research (IJSDR) that's, accuracy is dramatically improved when classifiers are applied to the classifier, Dataset instead of the information set. We use the Naïve Bayes Classifier here and extract the word using the word count algorithm. The error rate is incredibly low after we use the Bayesian Naïve Classifier. It will be said that Naïve Bayesian Classifier produces an improved result than Support Vector Machine.

REFERENCES:

[1]. Masurah Mohammad, Ali Selman “An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification”, IEEE,2015.

[2]. C. Bala Kumar, D. Ganesh Kumar “A Data Mining Approach on Various Classifiers in Email Spam Filtering”, IJRASET, May 2015

[3]. Vinod Patidar, Divakar Singh, Anju Singh "A Novel Technique of Email Classification for Spam Detection ", International Journal of Applied Information Systems (IJAIS), Volume 5 – No. 10, August 2013.

[4]. Cormack, Gordon. Smucker, Mark. Clarke, Charles "Efficient and effective spam filtering and re-ranking for large web datasets" Information Retrieval, Springer Netherlands. January 2011

[5]. Archit Mehta ,Raunakraj Patel “Email Classification using data Mining”, IJARCCE, 2011.

[6]. I. Androutsopoulos. (2000). Ling-Spam. Aueb.gr. Accessed: Oct. 11, 2019. [Online].

[7] X.L. Wang Learning to classify email: a survey 2005 International Conference on Machine Learning and Cybernetics (Vol. 9, pp. 5716-5719), IEEE (Aug 2005)

[8]. I. Androutsopoulos, V. Metsis, and G. Paliouras. (2006). The Enron-Spam Datasets. Accessed: Oct. 11, 2019. [Online].
[9]. R. Shams and R. E. Mercer, ‘‘Classifying spam emails using text and readability features,’’ in Proc. IEEE 13th Int. Conf. Data Mining, Dallas, TX, USA, Dec. 2013, pp. 657–666.
[10]. C.P. Lueg From spam filtering to information retrieval and back: seeking conceptual foundations for spam filtering Proc. Assoc. Inf. Sci. Technol., 42 (1) (2005)
[11]. G.V. Cormack Email spam filtering: a systematic review Found. Trends Inf. Retr., 1 (4) (2008), pp. 335-455
[12]. E.P. Sanz, J.M.G. Hidalgo, J.C.C. Pérez Email spam filtering Adv. Comput., 74 (2008), pp. 45-114
[13]. A. Bhowmick, S.M. Hazarika Machine Learning for E Mail Spam Filtering: Review, Techniques and Trends arXiv:1606.01042v1 [cs.LG] 3 Jun 2016 (2016), pp. 1-27
[14].C. Laorden, X. UgartPedrero, I. Santos, B. Sanz, J. Nieves, P.G. Bri ngas Study on the effectiveness of anomaly detection for spam filtering Inf. Sci., 277 (2014).