Optimizing Spam Filtering with Machine Learning

Project Description:

Over recent years, as the popularity of mobile phone devices has increased, Short Message Service (SMS) has grown into a multi-billion dollar industry. At the same time, reduction in the cost of messaging services has resulted in growth in unsolicited commercial advertisements (spams) being sent to mobile phones. Due to Spam SMS, Mobile service providers suffer from some sort of financial problems as well as it reduces calling time for users. Unfortunately, if the user accesses such Spam SMS they may face the problem of virus or malware. When SMS arrives at mobile it will disturb mobile user privacy and concentration. It may lead to frustration for the user. So Spam SMS is one of the major issues in the wireless communication world and it grows day by day.

To avoid such Spam SMS people use white and black list of numbers. But this technique is not adequate to completely avoid Spam SMS. To tackle this problem it is needful to use a smarter technique which correctly identifies Spam SMS. Natural language processing technique is useful for Spam SMS identification. It analyses text content and finds patterns which are used to identify Spam and Non-Spam SMS.

To accomplish this, we have to complete all the activities listed below,

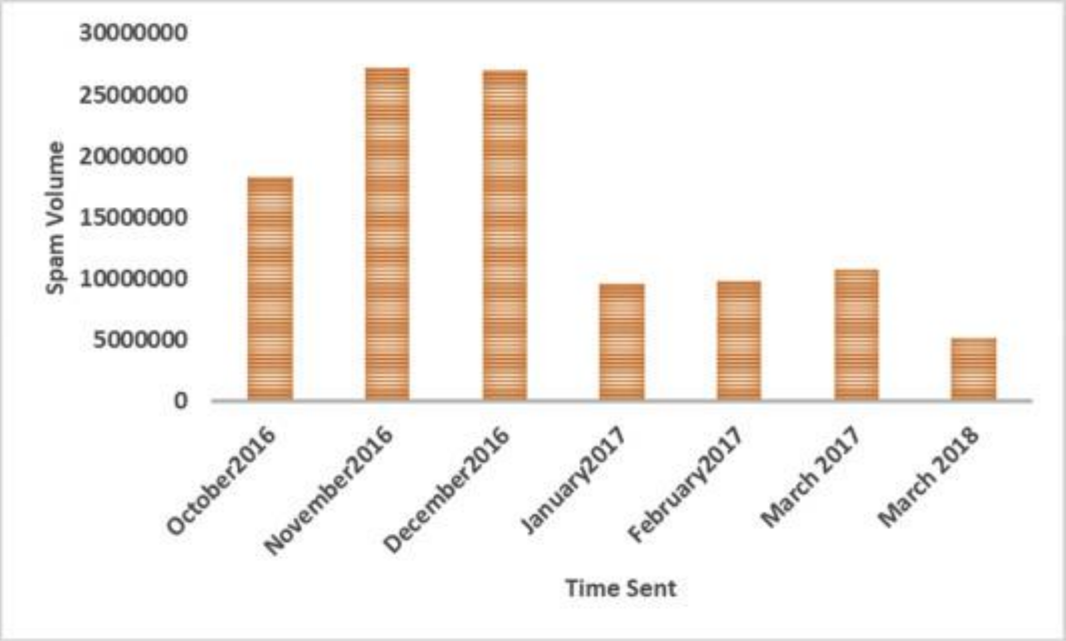
- Define Problem / Problem Understanding
  - Specify the business problem
  - Business requirements
  - Literature Survey
  - Social or Business Impact.

INTRODUCTION:

In recent times, unwanted commercial bulk emails called spam has become a huge problem on the internet. The person sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chatrooms, and viruses [1]. Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time [2]. The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the whole global email traffic [3]. Users who receive spam emails that they did not request find it very irritating. It is also resulted to untold financial loss to many users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails pretending to be from reputable companies with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers.

According to report from Kaspersky lab, in 2015, the volume of spam emails being sent reduced to a 12-year low. Spam email volume fell below 50% for the first time since 2003. In

June 2015, the volume of spam emails went down to 49.7% and in July 2015 the figures was further reduced to 46.4% according to anti-virus software developer Symantec. This decline was attributed to reduction in the number of major botnets responsible for sending spam emails in billions. Malicious spam email volume was reported to be constant in 2015. The figure of spam mails detected by Kaspersky Lab in 2015 was between 3 million and 6 million. Conversely, as the year was about to end, spam email volume escalated. Further report from Kaspersky Lab indicated that spam email messages having pernicious attachments such as malware, ransomware, malicious macros, and JavaScript started to increase in December 2015. That drift was sustained in 2016 and by March of that year spam email volume had quadrupled with respect to that witnessed in 2015. In March 2016, the volume of spam emails discovered by Kaspersky Lab is 22,890,956. By that time the volume of spam emails had skyrocketed to an average of 56.92% for the first quarter of 2016. Latest statistics shows that spam messages accounted for 56.87% of e-mail traffic worldwide and the most familiar types of spam emails were healthcare and dating spam. Spam results into unproductive use of resources on Simple Mail Transfer Protocol (SMTP) servers since they have to process a substantial volume of unsolicited emails [127]. The volume of spam emails containing malware and other malicious codes between the fourth quarter of 2016 and first quarter of 2018 is depicted in Fig. 1 below.



1 - SPECIFY THE BUSINESS PROBLEM:

This section discusses the research gaps and open research problems of the spam detection and filtration domain. In the future, experiments and models should be trained on real-life data rather than manually created datasets, because, in the various article, the models trained on artificial datasets perform very poorly on real-life data. Currently, supervised, unsupervised, and reinforcement learning algorithms are used for spam detection, but we can get higher accuracy and efficiency by using hybrid algorithms in the future. Feature extraction can be improved in the future by using deep learning for feature extraction. Using clustering techniques for spam filtering relevance feedback using dynamic updating can better cluster spam and ham. Along with machine learning, blockchain models and concepts can also be used for email spam detection in the future. Experts in linguistics and psycholinguistics can collaborate in the future for manual annotation of datasets, which will result in the development of effective and standard spam datasets with high dimensionality. In future, spam filters can be designed with faster processing and classification accuracy using Graphics Processing Units (GPUs) and Field Programmable Gate Arrays (FPGAs), which offer low energy consumption, flexibility, and real-time processing capabilities. Moreover, future research should concentrate on the availability of standard labeled datasets for researchers to train classifiers and the addition of more attributes to the dataset to improve the accuracy and reliability of spam detection models, such as the spammer’s IP address and the location. The following are some other future research directions and open research problems in the domain of spam detection.(i)Some studies considered header, subject of the email, and message body as a feature for spam classification. While these features are not enough for fully accurate results, manual feature selection and features should also be.(ii)Almost all researchers presented their results based on accuracy, precision, recall, etc., while the time complexity of machine learning models should be considered an evaluation metric.(iii)Some researchers show promising results in the process of feature extraction using a bag of words. They claim that the email header is as important for spam detection as the content of the body. So, deep feature extraction of the header line should be considered.(iv)Fault tolerance, self-learning, and quick response time can be better by using comprehensive feature engineering and an accurate preprocessing phase.(v)Deep learning models with dynamic updating of feature space are needed to implement for better spam classification. Most of the current filters cannot update their feature space.(vi)The security of spam detection and filtration system is needed for better accuracy and reliable results.(vii)The false positive rate of many models is still higher than required. It must be reduced to the smallest possible value.(viii)Few spam filters work on image spam detection and filtration. Expert spammers also use images for spam messages, so it should be considered in detecting spam.(ix)Real-time spam classification is much needed as most of the proposed models cannot work on real-time data.(x)Labeled data is one of the major issues in spam detection. There are a few new labeled and up-to-date datasets for this purpose.(xi)Multilingual spam detection is also a significant research area that can be explored for better spam detection systems. There is less work done on multilingual spam detection using deep learning techniques.(xii)Semisupervised and federated learning techniques can be used to enhance spam detection in various IoT and email frameworks.(xiii)A combination of linguistic features for the spam detection approach can also be explored.(xiv)The research community ignores the identification of spammers and spammer networks.(xv)Many researchers manually annotate data, using spam features that they think to be accurate. As a result, the evaluation results of the detection systems that they propose are doubted. The ideal solution for this problem has yet to be discovered.(xvi)There is a lack of a robust method of dealing with challenges regarding the spam filters' security. An attack of this nature can be a casual, exploratory, or targeted attack. The deep learning techniques with blockchain technology can be used for this purpose.

1.1 - CHALLENGES OF SPAM DETECTION

Some critical challenges faced by spam filters are discussed as follows:(i)The growing amount of data on the Internet with various new features is a big challenge for spam detection systems.(ii)Features' evaluation from several dimensions such as temporal, writing styles, semantic, and statistical ones is also challenging for spam filters.(iii)Most of the models are trained on balanced datasets, while self-learning models are not possible.(iv)Many spam detection models face adversarial machine learning attacks that will decrease their effectiveness. Adversaries can throw a variety of attacks during the training and testing of ML models. Adversaries can harm training data to cause a classifier to classify the data incorrectly (poisoning attack), create unfavorable samples during testing to evade detection (evasion attack), and obtain sensitive training data via a learning model (privacy attack)(v)Deep fake is another big challenge that is being faced by spam detection systems. To generate, modify, and style pictures and videos, neural network models such as GPT-2,3 and image generation models like BigGAN, StyleGAN, and CycleGAN are adopted. Deep fakes can be used to disseminate false information.

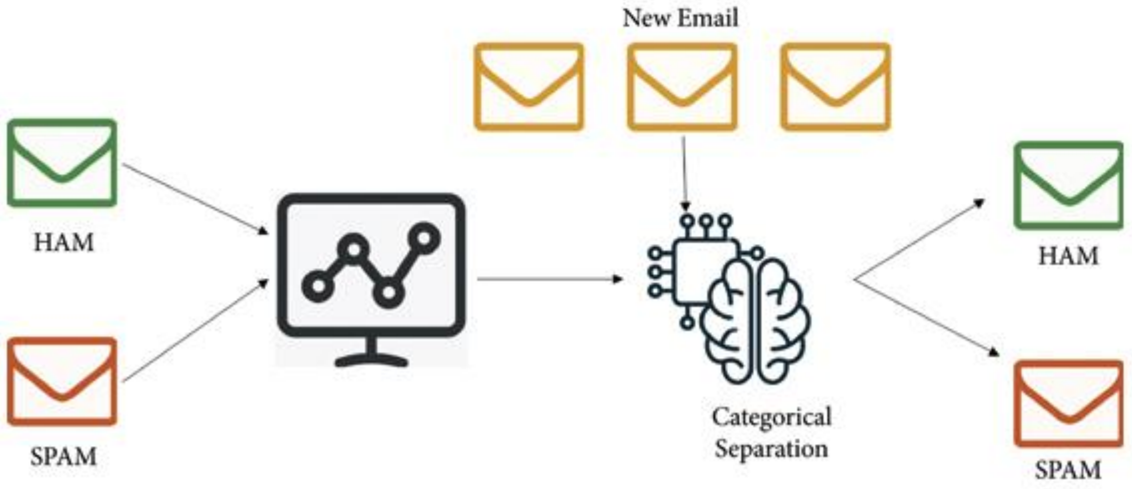
1.2 PROBLEM STATEMENT

A tight competition between filtering method and spammers is going on per day, as spammers began to use tricky methods to overcome the spam filters like using random sender addresses or append random characters at the beginning or end of mails subject line. There is a lack of machine learning focuses on the model development that can predict the activity. Spam is a waste of time to the user since they have to sort the unwanted junk mail and it consumed storage space and communication bandwidth. Rules in other existing must be constantly updated and maintained make it more burden to some user and it is hard to manually compare the accuracy of classified data.

2 – BUSINESS REQUIRMENTS:

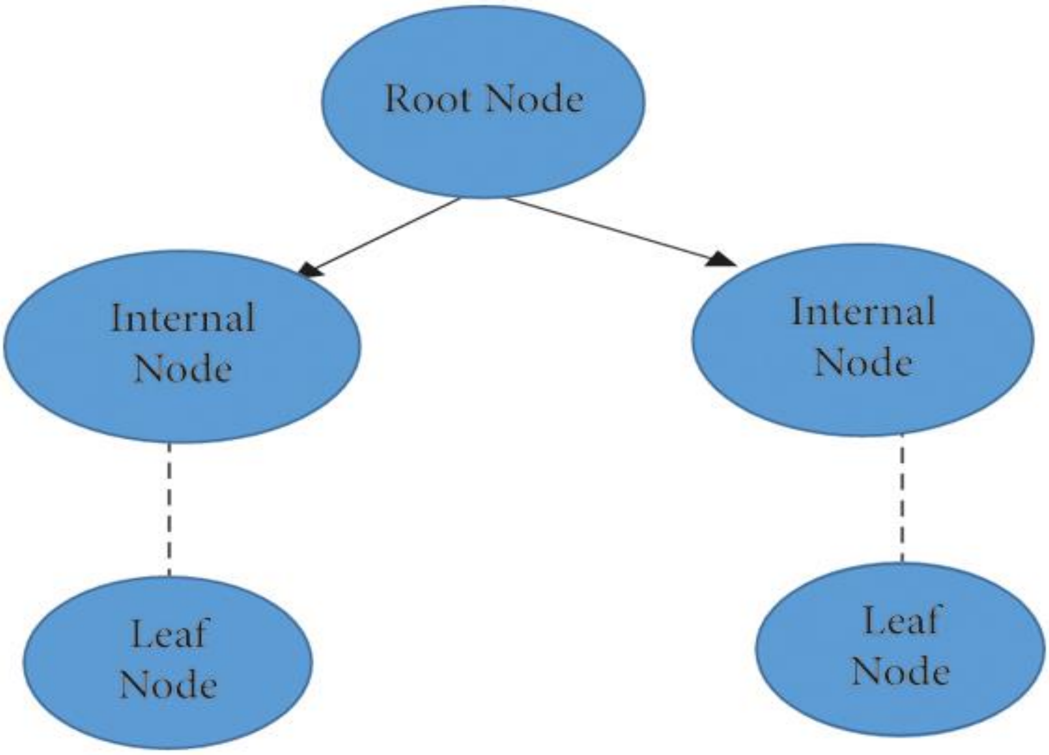
Supervised machine learning algorithms [18] are machine learning models that need labeled data. Initially, labeled training data is provided to these models for training, and after training models predict future events. In other words, these models begin with the analysis of an existing training dataset, and they generate a method to make predictions of success values. Upon proper training, the system can provide [38] the prediction on any new data related to the user's data at the training time. Furthermore, the learning algorithm accurately compares the output to the expected output and identifies errors to modify the model.

Supervised learning uses labeled data for training, and then it can predict the new data. This type of learning can be used in solving various problems, i.e., advertisement popularity, spam classification, face recognition, and object classification.



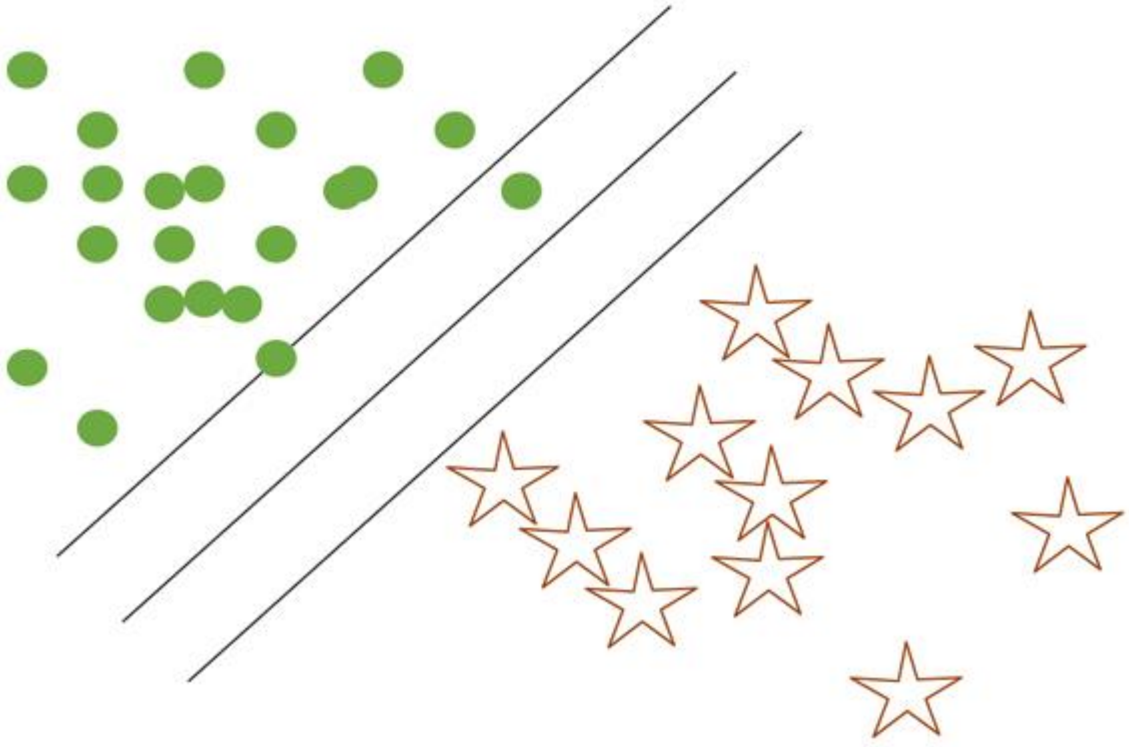
DECISION TREE CLASSIFIER

Decision tree classifier is a machine learning algorithm [39], which has been widely used since the last decade for classification. This algorithm applies a simple method of solving any problem of classification. A decision tree classifier is a collection of well-defined questions about test record attributes. Each time we get an answer, a follow up question is raised until a decision is not made on the record [40]. Tree-based decision algorithms define models that are constructed iteratively or recurrently based on the data provided. The decision tree-based algorithms goal is used to predict a target variable's value on a given set of input values. This algorithm uses a tree structure to solve classification and regression problems



SUPPORT VECTOR MACHINE (SVM)

The support vector machine (SVM) is an essential and valuable machine learning model [53]. SVM is a formally defined discriminative supervised learning classifier that takes labeled examples for training and gives a hyperplane as output, classifying new data [54]. A set of objects belonging to various class memberships are separated by decision planes. Figure 9 shows the classification concept of linear support vector machines. In the figure, some circles and stars are called objects. These objects can belong to any of two classes, i.e., the class of stars or dots. The isolated lines determine the choice of objects between green and brown objects. On the lower side of the plane, the objects are brown stars, and on the upper side of the plane all objects are green dots showing that two unique objects are classified into two different classes. If a new object black circle is given to the model, it will classify that circle into one of the classes according to the training examples provided in the training phase.



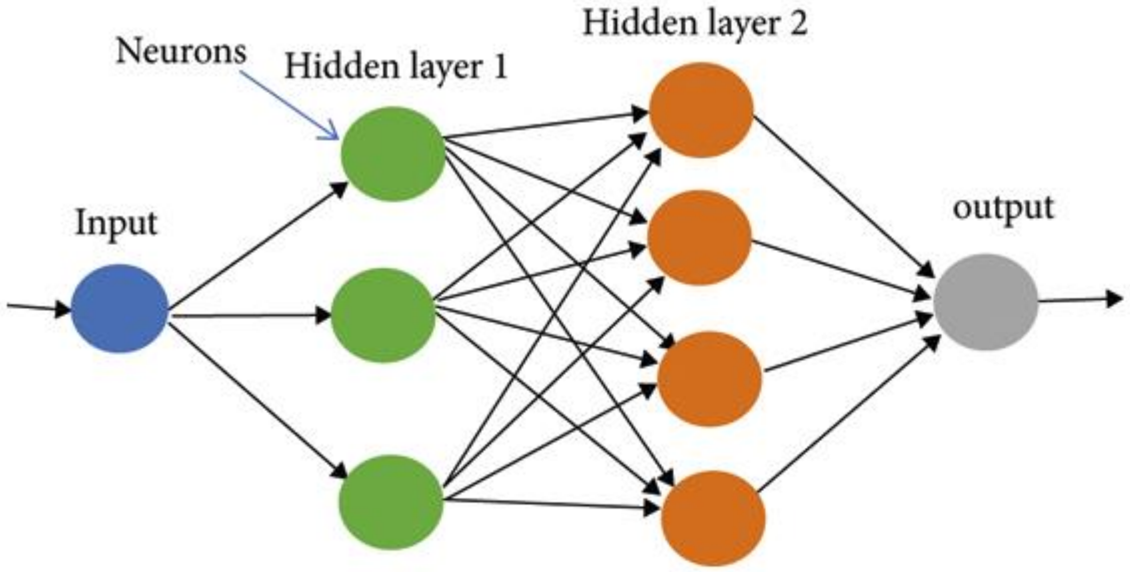
NAÏVE BAYES CLASSIFIER (NB)

The Naïve Bayes classifier [47] is based on the Bayes theorem. It assumes that the predictors are independent, which means that knowing the value of one attribute impacts any other attribute’s value. Naïve Bayes classifiers are easy to build because they do not require any iterative process and they perform very efficiently on large datasets with a handsome level of accuracy. Despite its simplicity, Naïve Bayes is known to have often outperformed other classification methods in various problems.

Kumar et al. [14] discussed email spam detection using various ML algorithms. Their article explores ML methods and how to implement them on datasets. The optimal algorithm for email spam detection with the highest precision and accuracy is identified from various ML algorithms. They concluded that the Multinomial Naïve Bayes algorithm produces the best results, but it has limitations due to class-conditional independence, which causes the machine to misclassify some inputs. Ensemble models come after Multinomial Naïve Bayes with the best and reliable results in this study. The proposed system in this study can only detect spam from the body of emails.

ARTIFICIAL NEURAL NETWORKS

An artificial neural network (ANN) is a computational model based on the functional aspects of biological neural networks, also known as the neural network (NN) [66]. Many sets of neurons are joined in a neural network, and information is interpreted using a computational approach connection. In most situations, an ANN is an adaptive system, which changes its structure depending on external or internal information flowing through the network during the learning phase. Current neural networks are nonlinear approaches to statistical data processing. These are commonly used when there are complex relationships between inputs and outputs or unusual performance patterns [6]. Figure 10 shows the basic structure of the neural network.



3 - LITERATURE SURVEY:

Email spam is nothing more than fake or unwanted bulk mails sent via any account or an automated system. Spam emails are increasing day by day, and it has become a common problem over the last decade. Email IDs receiving spam emails are typically collected through spambots (a computerized application that crawls email addresses across the Internet). The applications of machine learning have been playing a vital role in the detection of spam emails. It has various models and techniques that researchers are using to develop novel spam detection and filtering models [13]. Kaur and Verma [14] present a survey on email spam detection using a supervised approach with feature selection. They discuss the knowledge discovery process for spam detection systems. They also elaborate various techniques and tools proposed for spam detection. The choice of features based on N-Gram is also addressed in this survey. N-Gram [15, 16] is a predictive-based algorithm used to predict the probability of the next word occurrence after finding $N-1$ terms in a sentence or text corpus. N-Gram uses probability-based techniques for the next word prediction. They compare various machine learning (multilayer perceptron neural network support vector machine, Naïve Bayes) and nonmachine learning (Signatures, Blacklist and Whitelist, and mail header checking) approaches for email spam detection.

Saleh et al. [17] present a survey on intelligent spam email detection. They discuss various security risks of emails, especially spam emails, the scope of spam analysis, and different machine learning and nonmachine learning techniques for spam detection and filtering. They conclude that there is high adoption of supervised learning [18] algorithms for email spam detection. They state that the high usage of supervised learning is the accuracy and consistency of supervised techniques. They also discussed multialgorithm frameworks and found that multialgorithm frameworks are more efficient than a single algorithm. They found that nearly all research work that uses the content of emails for the identification spam, particularly phishing emails, depends on word-based classification or clustering systems.

Blanzieri and Bryl [2, 19] describe a list of learning-based email spam filtering approaches. In this paper, they addressed the spam problems and provided a review of learning-based spam filtering. They explain various features of spam emails. In this study, effects of spam emails on different domains were discussed. Various economic and ethical issues of spam are also discussed in this study. The antispam approach that is common and learning-based filtering is well developed. The commonly used filters are based on different classification techniques applied to various components of email messages. This study suggests that the Naïve Bayes classifier holds a particular position amongst multiple learning algorithms used for spam filtering. With splendid pace and simplicity, it gives high precision results.

Bhuiyan et al. [20] present a review of current email spam filtering approaches. They summarize multiple spam filtering approaches and sum up the accuracy on various parameters of different proposed systems by analyzing numerous processes. They discuss that all the existing methods are efficient for filtering spam emails. Some have successful results, and others are attempting to incorporate other ways to boost their accuracy performance. Although they are all successful, they still have some issues in spam filtering methods, which is the primary concern for researchers. They are trying to create a next-generation spam filtering mechanism to understand large numbers of multimedia data and filter spam emails. They conclude that most email spam filtering is done by utilizing Naïve Bayes and the SVM algorithm. To test the spam filtration models, these models can be trained on different datasets, such as “ECML” and UCI dataset [21].

Ferrag et al. [13] presented a review of deep learning algorithms of intrusion detection systems and spam detection datasets. They discussed various detection systems based on deep learning models and evaluated the effectiveness of those models. They examined 35 well-known cyber dataset by dividing them into seven categories. These categories include Internet traffic-based, network traffic-based, Interanet traffic-based, electrical network-based, virtual private network-based, andriod apps-based, IoT traffic-based, and Internet connected device-based datasets. They conclude that deep learning models can perform better than traditional machine learning and lexicon models for intrusion and spam detection.

Vyas et al. [22] present a review on supervised machine learning strategies for filtering spam emails. They concluded that the Naïve Bayes method provides faster results and decent precision over all other methods (except SVM and ID3) from all the techniques discussed. SVM and ID3 offer greater precision than Naïve Bayes but take much longer time to construct a system. There is a trade-off between timing and precision. They conclude that selecting the learning algorithm heavily depends on the situation and the required accuracy and time. They state that all parts of the email should be considered in the future to create a more robust spam filtering framework.

This survey paper discusses three main types of machine learning that can be used for spam filtering. We review various papers, the proposed techniques, and discuss challenges to spam detection and filtration systems. This article also focuses on the advantages and disadvantages of the proposed techniques for spam detection and filtration that is never reviewed in the past.

LITERATURE REVIEW

This chapter discusses about the literature review for machine learning classifier that being used in previous researches and projects. It is not about information gathering but it summarizes the prior research that related to this project. It involves the process of searching, reading, analysing, summarising and evaluating the reading materials based on the project.
Literature reviews on machine learning topic have shown that most spam filtering and detection techniques need to be trained and updated from time to time. Rules also need to be set for spam filtering to start working. So eventually it become burdensome to the user.

4 – SOCIAL OR BUSINESS IMPACT:

Spam can have a negative impact on any organisation, from decreased productivity and resources to overwhelming employees with unwanted messages. To stay competitive, businesses must proactively protect themselves against spam online via email.

Spam is a blanket term for unsolicited and often unwanted emails. It affects businesses by clogging up inboxes and causing recipients to delete emails without reading them, potentially missing out on significant opportunities. Moreover, spam can lead to decreased customer loyalty among consumers who expect companies they engage with to protect their data and privacy.

Fortunately, there are solutions such as spam filters that can help reduce the amount of these messages businesses receive – but some spammers continue to find ways around these filters and send damaging or offensive content into networks. Therefore, businesses must stay vigilant in managing the type of content on their digital channels. Spam represents over 50% of email traffic, which is the most intrusive method cyber criminals use to introduce malware to corporate systems.

This can greatly interrupt workflow and decrease business efficiency. With the help of Cybersecurity Training, your business can stay on top of phishing emails by training your team from within to recognise these attacks first-hand.

THE RISKS ASSOCIATED WITH SPAM

Spam emails can cause individuals and businesses many problems if they are opened and responded to; this shows the importance of Cybersecurity Training. By opening this mail, the reader confirms that their email address is active and valid and could download malware onto their device by clicking malicious links.

Additionally, responding can result in personal information, such as banking details, being stolen or money being stolen by unsuspecting victims. It’s recommended that people never open or respond to spam emails, even if the sender looks legitimate. If you need clarification, contacting the sender through another channel is best to verify the message before further action.

Finally, one of the most prominent risks of spam is DDoS attacks. DDoS stands for Distributed Denial of Service. This is when a hijacker gaining information through spam overloads your network bandwidth. This makes it, so you cannot use your network and cause long periods of downtime. This can damage a business’s image and decrease efficiency.

To prevent this, you should implement email filtering software for your business and train your staff from within with Cybersecurity Training so that even if a spam email goes through, your staff is aware of this phishing attack and doesn’t click any harmful links.

HOW TO IDENTIFY AND PREVENT SPAM

Guarding your mailbox against potential scams can be a manageable task with the help of Cybersecurity Training. Common signs to look for in determining if a scam is trying to reach your mailbox include offers that seem too good to be true, requests for large sums of money, or communications from someone acting suspiciously. If you receive any of these communications, you should talk to family or friends and research the offer before responding.

Additionally, it pays off to be sceptical about emails, even when they appear authentic; double-check before clicking anything, as malicious links can lead to disastrous results. Lastly, throw out any solicitations you receive without opening them, as these are likely some scams. These simple steps will help you protect yourself against scammers and give you peace of mind.

HOW TO REDUCE SPAM IN YOUR MAILBOX

Avoiding spam can be tiresome and lengthy, but with a few simple best practices, you can reduce the spam in your inbox. First and foremost, use trusted antivirus software to ensure that malicious software doesn't enter your computer, which is often how spammers get access to your emails. Secondly, never give out your email address unless necessary; if possible, create an alternative dedicated account for activities like online shopping that require sharing it.

Lastly, though this might seem obvious, make sure not to click on any shady links or advertisements embedded inside an email you receive – these could be from malicious senders looking for personal information. With these few steps, you're well on your way to reducing the amount of junk mail flooding your inbox. This is just the first line of defence. With the help of anti-spam filtering software and Cybersecurity Training, this should add another layer of security to keep your data and computer safe.

. It affects productivity in your company

Spam is related to the loss of productivity in your company because it occupies your team with an unnecessary task. It's true that it might not take long to open a mailbox and delete all spam.

Nonetheless, spam is a waste of time and a distraction for your employees, who could be spending energy with more productive activities.

2. It affects your business services and makes the company more vulnerable

It costs almost nothing for spammers to send millions of emails. The problem for your business is a mass attack.

If your business isn't ready to handle a large number of messages, many of your services may be disrupted.

In addition to losing new business because the company's communication has been hampered, a mass attack often leaves your business vulnerable to other threats and cyberattacks.

3. It contains malware, such as ransomware and spyware

Today spam is one of the most widely used vectors for the spread of threats, including malware.

Those seemingly harmless links and attachments can pose a real threat to your business, hiding ransomware, spyware, and trojans, which allow the attacker to gain access to the computer and then to the entire company's network.

By the way, according to a Verizon report, about 94% of breaches involving malware occur through the use of malicious emails.

4. It contains phishing and spoofing threats

Spam is widely used for spoofing and phishing scams, which may also be related to malware propagation.

Phishing and spoofing are used when a cybercriminal creates a fake website to steal access credentials with the intention of hacking into your business network or gaining access to confidential information.

Imagine that the attacker may attempt to induce an employee to make a payment for an invoice that doesn't exist or to provide the system's access credentials.

5. It can cause legal problems

In an extreme case, depending on the type of spam, such as pornography spam, for example, your company may face legal problems due to misuse of email for illegal activity.

In addition, it may be that a cybercriminal hijacks and uses your company's domain to spread spam, which may have implications under some specific law.

These are extreme cases, but they deserve attention.

CONCLUSION:

In this study, we reviewed machine learning approaches and their application to the field of spam filtering. A review of the state of the art algorithms been applied for classification of messages as either spam or ham is provided. The attempts made by different researchers to solving the problem of spam through the use of machine learning classifiers was discussed. The evolution of spam messages over the years to evade filters was examined. The basic architecture of email spam filter and the processes involved in filtering spam emails were looked into. The paper surveyed some of the publicly available datasets and performance metrics that can be used to measure the effectiveness of any spam filter. The challenges of the machine learning algorithms in efficiently handling the menace of spam was pointed out and comparative studies of the machine learning technics available in literature was done. We also revealed some open research problems associated with spam filters. In general, the figure and volume of literature we reviewed shows that significant progress have been made and will still be made in this field. Having discussed the open problems in spam filtering, further research to enhance the effectiveness of spam filters need to be done. This will make the development of spam filters to continue to be an active research field for academician and industry practitioners researching machine learning techniques for effective spam filtering.

Our NM team members;

A.AGASTIN

J.VISHVA

M.NAVEEN KUMAR

O.V.KARTHI GIRISH

A.MADHAVAN