FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

Department of Computer Engineering

Course, Subject & Experiment Details

Practical No:	**1**
Title:	**IEEE – Email Spam Filtering**
Name of the Student:	**Warren Fernandes**
Roll No:	**8940**
Date of Performance:	**22/04/2022**
Date of Submission:	**22/04/2022**

Evaluation:

Sr. No.	Rubric	Grade
1	On time submission/completion (2)	
2	Preparedness (2)	
3	Skill (4)	
4	Output (2)	

Signature of the Teacher

EMAIL SPAM FILTERING

Abstract -
The path finder application helps to find the path between two places within limited defined area, beneficial for those who are new to an organization as they do not know the path inside the organization.
Artificial intelligence assistants is being implemented to find the effective path within the organization.
Keywords - artificial intelligence, automation, voice assistant, shortest path

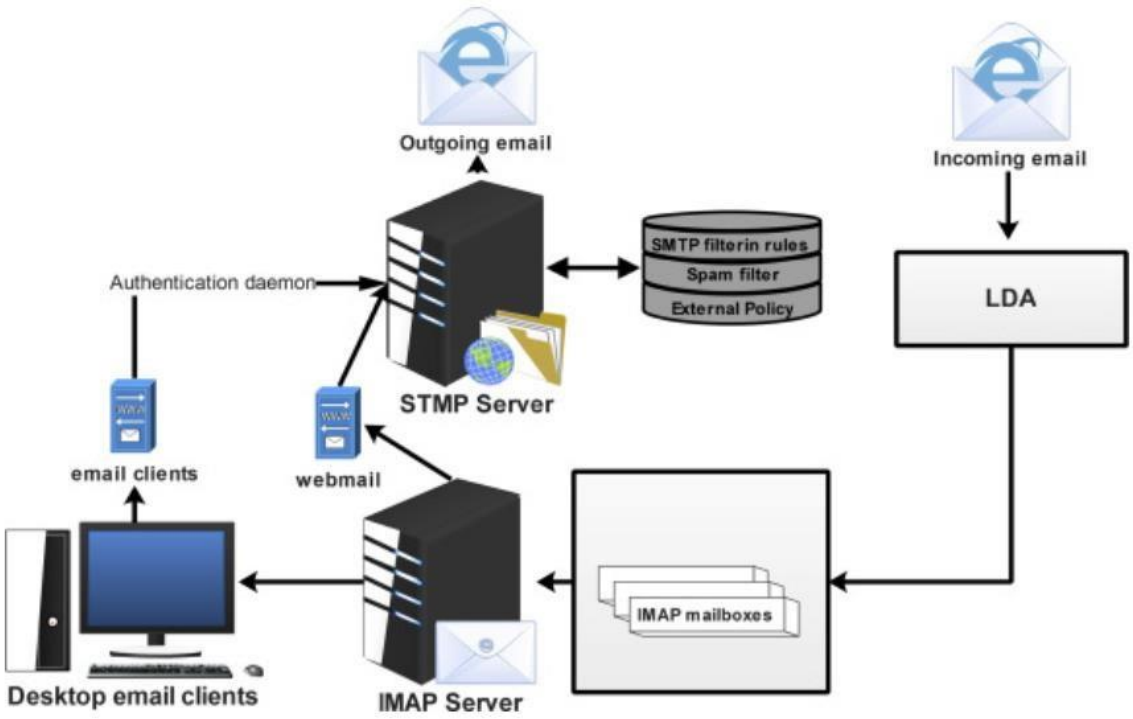
Introduction -
Different spam filtering formulas have been employed by Gmail, Outlook.com and Yahoo Mail to deliver only the valid emails to their users and filter out the illegitimate messages. Conversely, these filters also sometimes erroneously block authentic messages. It has been reported that about 20 percent of authorization based emails usually fail to get to the inbox of the expected recipient. The email providers have designed various mechanisms for use in email anti-spam filter to curtail the dangers posed by phishing, email-borne malware and ransomware to email users. The mechanisms are used to decide the risk level of each incoming email. Examples of such mechanisms include satisfactory spam limits, sender policy frameworks, whitelists and blacklists, and recipient verification tools. These mechanisms can be used by single or multiple users. When the satisfactory spam thresholds is too low it can lead to more spam evading the spam filter and entering the users' inboxes. Meanwhile having a very high threshold can lead to some important emails being isolated unless the administrator redirects them.

Description –
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. 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. The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the whole global email traffic . 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.

The necessary stages that must be observed in the mining of data from an email message can be categorised into the following:

1. Pre-processing: This is the first stage that is executed whenever an incoming mail is received. This step consists of tokenization.
2. Tokenization: This is a process that removes the words in the body of an email. It also transforms a message to its meaningful parts. It takes the email and divides it into a sequence of representative symbols called tokens.
3. Feature selection: Sequel to the pre-processing stage is the feature selection phase. Feature selection a kind of reduction in the measure of spatial coverage that effectively exemplifies fascinating fragments of email message as a compressed feature vector.



Problems In Existing Systems -

1. Failure of many spam filters to reduce their false positive rate.

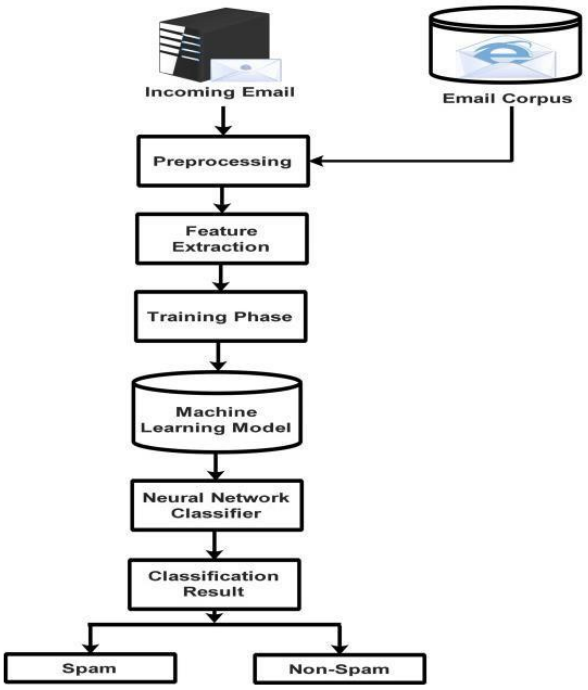
2. The inability of the current spam filtering techniques to effectively deal with the concept drift phenomenon.
3. Lack of effective strategy to handle the threats to the security of the spam filters. Such an attack can be causative or exploratory, targeted or indiscriminate attack.

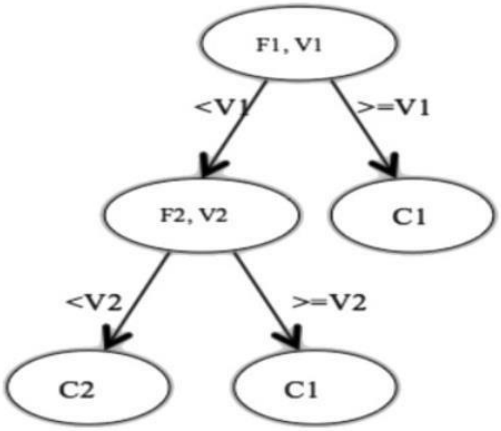
Discussions -

Machine learning algorithms have been extensively applied in the field of spam filtering. Substantial work have been done to improve the effectiveness of spam filters for classifying emails as either ham (valid messages) or spam (unwanted messages) by means of ML classifiers. They have the ability to recognise distinctive characteristics of the contents of emails. Many significant work have been done in the field of spam filtering using techniques that does not possess the ability to adapt to different conditions; and on problems that are exclusive to some fields e.g. identifying messages that are hidden inside a stego image.

Implementation –

The Bayesian classification exemplifies a supervised learning technique and at the same time a statistical technique for classification. It acts as a fundamental probabilistic model and let us seize ambiguity about the model in an ethical way by influencing the probabilities of the results. It is used to provide solution to analytical and predictive problems. Bayesian classification is named after Thomas Bayes (1702–1761), who proposed the algorithm. The classification offers practical learning algorithms and previous knowledge and experimental data can be merged. Bayesian Classification offers a beneficial viewpoint for comprehending and appraising several learning algorithms.





By partitioning the email dataset in relation to least entropy, the resultant email dataset has the highest information gain and so impurity (emails contain both spam and ham) of the dataset is reduced. The dataset can be tested using the decision tree algorithm after the tree is created from the training email dataset. The email dataset being tested undergo some processing in the tree using some predefined rules pending the time it will get to a leaf node. The label in the leaf node is then assigned to the tested data.

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. 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.

Research papers referred -

https://ieeexplore.ieee.org/abstract/document/7530239
https://link.springer.com/chapter/10.1007/978-3-642-13059-5_6
https://www.sciencedirect.com/science/article/pii/S240584401835340 4