Intelligent Spam Mail Detection System

Taher Jodiawala

Department of IT Engineering
University of Mumbai
Mumbai, Maharashtra, India
taher.jodiawala_19@sakec.ac.in

Manav Bhagia

Department of IT Engineering
University of Mumbai
Mumbai, Maharashtra, India
manav.bhagia_19@sakec.ac.in

Prateek Duhoon

Department of IT Engineering
University of Mumbai
Mumbai, Maharashtra, India
prateek.duhoon_19@sakec.ac.in

Shubhay Islaniya

Department of IT Engineering
University of Mumbai
Mumbai, Maharashtra, India
shubhay.islaniya_19@sakec.ac.in

Nivedeeta Mukherjee

Department of IT Engineering
University of Mumbai
Mumbai, Maharashtra, India
nivedeeta.mukherjee@sakec.ac.in

Nutan Dolzake

Department of IT Engineering
University of Mumbai
Mumbai, Maharashtra, India
nutan.dolzake@sakec.ac.in

---***---

***Abstract* – Emails are frequently used by individuals for professional and personal use. Many individuals possess more than one email id often provided by the organizations they are working with for professional use.[1] This indicates that multiple emails can be created and attackers make use of fake profile to con people by possessing as a genuine person from a legitimate organization. This is known as Email Phishing which is a popular cyber security attack used by attackers to gain sensitive information from users.[4] Nowadays, anyone can send an email to any organization or individual. This provides a golden opportunity to either send spam or malicious emails.[5] The goal of this paper is to identify these spam mails by using machine learning, which through its mechanisms, allows models to analyze massive amounts of complex data with the help of various algorithms and alert the user about suspicious and possibly spam mails.**

Keywords- Email, Spam, Phishing, Machine Learning, Accuracy

I. INTRODUCTION

Email Security Systems are essential security software/tools that are used mainly for protection against malicious email activities. Data privacy has become a major issue when communicating via email.[11] The user desires data secrecy and integrity, as well as a secure network through which data can be transferred. There are many dangerous activities such as phishing and virus, which infects our data and causes the system to behave inappropriately or abnormally or attack the system's functionality.[5] One of the major issues is that personal data of an organization's employees may contain extremely sensitive information, or trade secrets which can be leaked due to breaches is not an easy task.[9] There are many anti-virus products on the market for e-mail system security, but today's attackers have a wide range of unusual skills at their disposal, allowing them to change the virus's existing code and thus compromise the system's security. A security system paired with new-age technology like AI and machine learning can make it a lot more effective and efficient.[5] There is an ever-increasing need of spam detection systems as email-borne attacks are also evolving over time; email is a common social engineering channel that is widely used by scammers, hackers, and others. As emails can be rapidly sent to many people it becomes a game of probability of some victim being scammed or falling prey to such malicious activity.[10] It is not just the people who are technologically illiterate that fall to such attacks, individuals accessing emails on the daily basis can also not realize when they are being targeted by spam mails. It is the need of the hour to not only identify spam mails but also alert the user about the same. It has been observed that spam mails rely on social engineering more than the technical aspect of emails.[9]

Phishing attacks can not just be the usual lottery scams, attackers nowadays prepare a lot of information about their victims and customize their e-mails for them accordingly.[7] Earlier the usual observation regarding spam mails was that they were used for targeted advertising or just simple advertising, but nowadays attackers disguise themselves under the fake banner of a known organization in attempts to direct the user to a malicious or infected website.

II. LITERATURE REVIEW

A. Analysis

To get a better understanding of the problem at hand and to also analyse the working of current applications of the same domain, the group read and understood some literature papers. These papers not only helped the group in better understanding of the problem but also highlighted certain missed aspects.

Ref No	Paper	Algorithm Used	ACCURACY	HIGHEST ACCURACY ALGORITHM	DATASET	Paper Explanation
[1]	Efficient Email Phishing Detection Using Machine Learning [1]	LOGISTIC MODEL TREE-LMT	96.77%	LOGISTIC MODEL TREE-LMT	PHISH TANK()	Detection Of Phishing Emails
		MULTILAYER PERCEPTION-LMP	95.87%			
		DECISION TREE-J48	96.92%			
[2]	A Comparative Approach to Naive Bayes Classifier and Support Vector Machine for Email Spam Classification [2]	SUPPORT VECTOR MACHINE	93.50%	SUPPORT VECTOR MACHINE	Enron corpus	Detection Of Spam Or Legitimate Emails
		NAÏVE BAYES CLASSIFIER	92%			
[3]	Email Spam Detection Using Machine Learning Algorithms [3]	Support Vector Classifier	90%	NAÏVE BAYES CLASSIFIER	spam email data set from - Kaggle	Detection Of Spam Or Legitimate Emails
		K-Nearest Neighbour	88.75%			
		Naïve Bayes	95.25%			
		Decision Tree	94.25%			
		Random Forest	91.50%			
		AdaBoost Classifier	94.50%			
		Bagging Classifier	94.25%			
[4]	Detection of Phishing Emails using Machine Learning and Deep Learning [4]	logistic regression	99.80%	Random Forest		Detection Of Phishing Emails along with FLASK python application
		random forest				
		XG boosting				
[5]	Applying machine learning and natural language processing to detect phishing email [5]	Graph convolutional network (GCN)	98.20%	PVDBOW	Fraud Dataset 2010	Detection of Phishing Emails

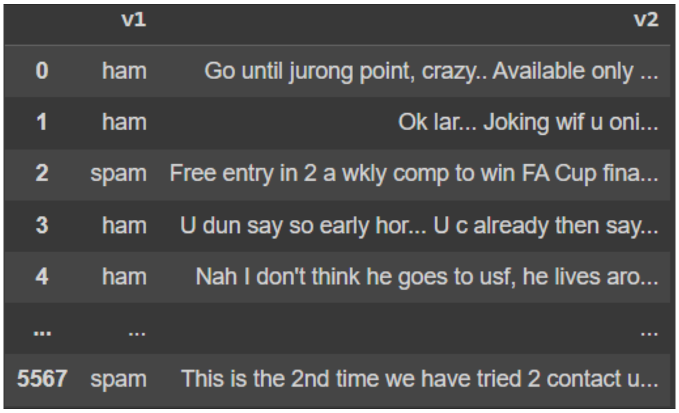
III. PROPOSED MODEL

A client is a person who can send or receive an email via the Internet or email network. Spam detection at the client level provides a multitude of rules and mechanisms to ensure secure communication transmission between individuals and organisations. A client must deploy numerous existing frameworks on his or her system for data transmission. These systems communicate with client mail agents in order to filter the client's mailbox by composing, accepting, and managing incoming emails.

IV. METHODOLOGY

A. Dataset

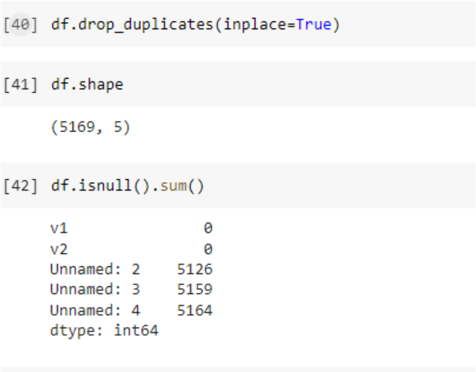
For the working of this model, the "Spam.csv" dataset from Kaggle has been used which has entries of roughly around 5500 consisting of 2 columns namely spam/ham detection column and text columns.



	v1	v2
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...

B. Data Preprocessing

In data pre-processing we perform data cleaning by learning and finding out about null entries. We also found about repeated entries to make our dataset cleaner. Further we have visualized the data and to get a better view of the dataset.



```
[40] df.drop_duplicates(inplace=True)

[41] df.shape

     (5169, 5)

[42] df.isnull().sum()

     v1               0
     v2               0
     Unnamed: 2    5126
     Unnamed: 3    5159
     Unnamed: 4    5164
     dtype: int64
```

C. NLTK library

It is the platform that can help us work with human language, working with fundamentals of writing programs, working with the corpus (paragraph, sentences), categorizing text, analysing linguistic structure, and more.[7] Stopwords, are the words which have no significance in giving the sentence a meaning but just help in forming it so that they make sense. To make data processing easier we eradicate them.

For example: "Yay!! You have won a gift hamper worth 7000." As you can analyze that 'you, have, a' are of no significance they are just adding weight-age to our data.

Using NLTK toolkit which is used to pre-process text which is in human readable format and mostly unorganized, to make it eligible for analysing.

D. Naïve Bayes

Naive Bayes is based on Bayes' Theorem Formula with a premise of independence among predictors.[12] Given a Hypothesis A and evidence B, Bayes' Theorem calculator states that the relationship between the probability of Hypothesis before getting the evidence P(A) and the probability of the hypothesis after getting the evidence P(A|B) is:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Here:

- A, B = events
- P(A|B) = probability of A given B is true
- P(B|A) = probability of B given A is true
- P(A), P(B) = the independent probabilities of A and B

This theorem, as explained in one of our previous articles, is mainly used for classification techniques in data analytics. The Naive Bayes theory calculator is essential for detecting email spam. [3] Naive Bayes is a simple probability strategy that assumes each characteristic of the model is independent of the others. [12] In the context of the spam filter, we assume that each word in the message is distinct from every other word and we tally the words without taking context into account. Using the current collection of terms, our classification algorithm generates probability for whether a message is spam or not. The Bayes formula is used to determine the likelihood, and the formula's individual components are based on the word frequencies over the entire set of messages. One of the main uses of machine learning in today's innerwebs is spam detection.[9] Almost all of the major email service providers have built-in spam detection systems that categorise such material as "Junk Mail" when it is received.

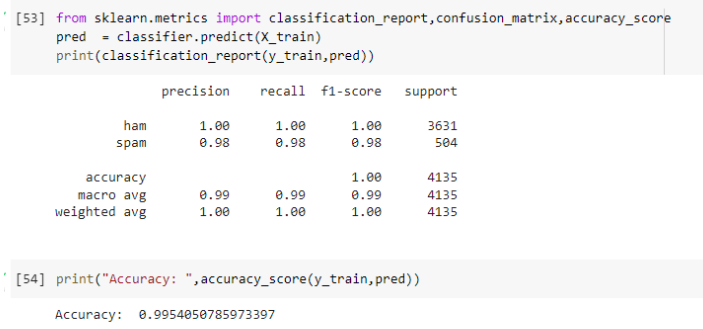
In this mission, we'll use the Naive Bayes method to build a model that, depending on the model's training data, can determine whether a dataset of SMS texts are spam or not. It's critical to have some sense of what a spammy SMS message might resemble. These texts are typically written with phrases like "free," "win," "winner," "cash," "prize," and the like since they are intended to attract your attention and, in a manner, persuade you to open them.[1] Additionally, exclamation points and all-caps writing are common features of spam texts. We want to train a model to recognise spam texts for us because, to the recipient, they are typically quite obvious.

Since messages can only be categorised as "Spam" or "Not Spam" and nothing else, identifying spam messages is a binary classification problem. Additionally, since we will be providing the model a labelled dataset that it can use to learn from and make future predictions, this is a supervised learning problem.

Confusion matrix

Another statistic that is frequently used to gauge how well a classification system is performing is the confusion matrix. Despite its name, the confusion matrix's nomenclature can be somewhat perplexing, but the matrix itself is easy to understand.

- Recall: The ability of a model to find all the relevant cases within a data set. Mathematically, we define recall as the number of true positives divided by the number of true positives plus the number of false negatives.
- Precision: The ability of a classification model to identify only the relevant data points. Mathematically specifies the number of true positives divided by the number of true positives and the number of false positives.
- The F1 score is a combination of precision and recall for a particular positive class. The F1 score can be interpreted as a weighted average of precision and recall. The highest F1 score is 1 and the lowest is 0.
- Support is the actual number of occurrences of the class in the specified record. Disproportionate support in the training data may indicate structural weaknesses in the scores reported by the classifier, and may indicate the need for stratified sampling or recalibration.



```
[53] from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
     pred  = classifier.predict(X_train)
     print(classification_report(y_train,pred))
```

```
              precision    recall  f1-score   support

         ham       1.00      1.00      1.00      3631
        spam       0.98      0.98      0.98       504

    accuracy                           1.00      4135
   macro avg       0.99      0.99      0.99      4135
weighted avg       1.00      1.00      1.00      4135
```

```
[54] print("Accuracy: ",accuracy_score(y_train,pred))
```

```
Accuracy:  0.9954050785973397
```

PDF SCANNING AND VIRUS LINK SCANNING

As we can made some additional improvements to our on-going project as well as try to make the emails received safer and more made sure they are less malicious and safe for the user to use and interact with. For the purpose of attachments that are attached to the emails. The API used to Scan is VirusTotal API which uses well known python library that contains latest definitions and schematics of the viruses and malwares existing in the internet space and market

The VirusTotal API offers several functionalities such as file uploading and scanning, URL submission and scanning, accessing completed scan reports, and automatic commenting on URLs and samples, all without requiring the use of the HTML website interface.

Real-time updates

At VirusTotal, malware signatures are regularly updated by leading antivirus companies to provide our users with the most up-to-date signature sets. Whenever a contributor blacklists a URL, our service immediately reflects the changes in the user-facing verdicts.

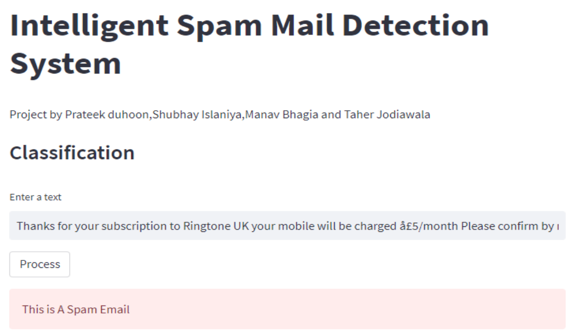
Phishing url Scanning

url phishing is an integral part of email phishing. Emails are used to redirect users to fraudulent websites and coerce them into divulging sensitive information such as sharing login credentials or transfer of funds. A phishing url dataset was pre-processed and two machine learning models were trained namely Logistic Regression and Multinomial Naïve Bayes with accuracy of 96% and 95% respectively. Logistic regression provided a better accuracy as compared to Multinomial Naïve Bayes.

Using pickle a python library we dump the machine learning model to convert it into binary format to publish it on web page. Python application was coded to publish the result on the web page using FastAPI which is a web framework for building RESTful APIs in python. The web page consisted of an input in which urls can be predicted to check whether it is a phishing url or not.

V. RESULT

Spam Mail:



Intelligent Spam Mail Detection System

Project by Prateek duhoon,Shubhay Islaniya,Manav Bhagia and Taher Jodiawala

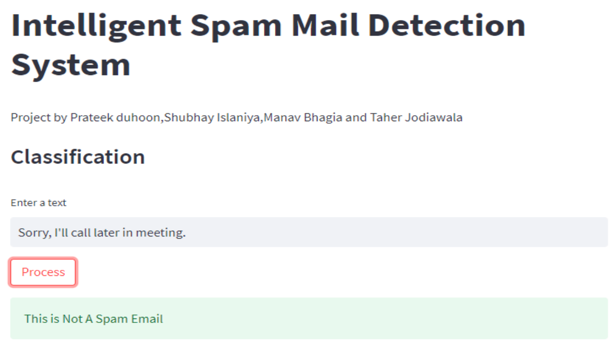
Classification

Enter a text

Thanks for your subscription to Ringtone UK your mobile will be charged â£5/month Please confirm by

This is A Spam Email

Ham Mail:



Intelligent Spam Mail Detection System

Project by Prateek duhoon,Shubhay Islaniya,Manav Bhagia and Taher Jodiawala

Classification

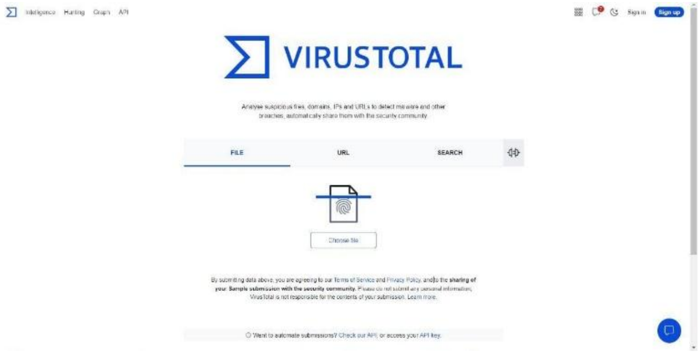
Enter a text

Sorry, I'll call later in meeting.

This is Not A Spam Email

Pdf Scanner:

VirusTotal provides more than just a binary detection result for submitted files, as it also displays the detection labels of each antivirus engine used in the scan, such as I-Worm. Allaple. gen. Similarly, for URL scanning, it offers a detailed analysis that can distinguish between various types of harmful sites, including malware, phishing, and suspicious sites. Some engines provide further information, such as identifying which botnet a URL belongs to or which brand a phishing site is targeting.



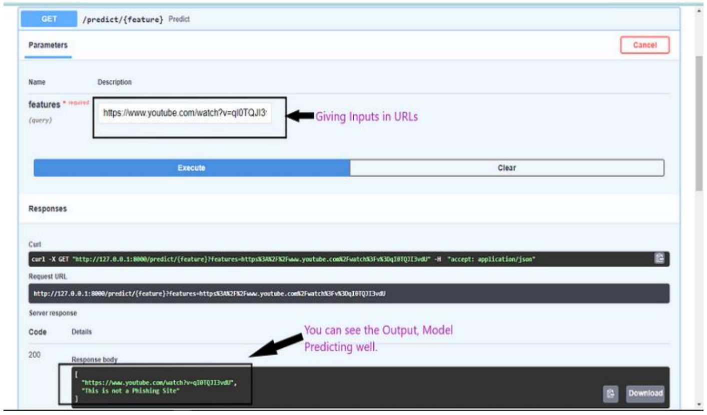
Analyse suspicious files, domains, IPs and URLs to detect malware and other breaches, automatically share them with the security community

By submitting data above, you are agreeing to our Terms of Service and Privacy Policy, and to the sharing of your Sample submission with the security community. Please do not submit any personal information; VirusTotal is not responsible for the contents of your submission. Learn more.

Want to automate submissions? Check our API, or access your API key.

Phishing Url:

Phishing Site



features * required (query): https://www.youtube.com/watch?v=ql0TQJI3

Giving Inputs in URLs

Curl

```
curl -X GET "http://127.0.0.1:8000/predict/{feature}?features=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DqI0TQJI3vdU" -H  "accept: application/json"
```

Request URL

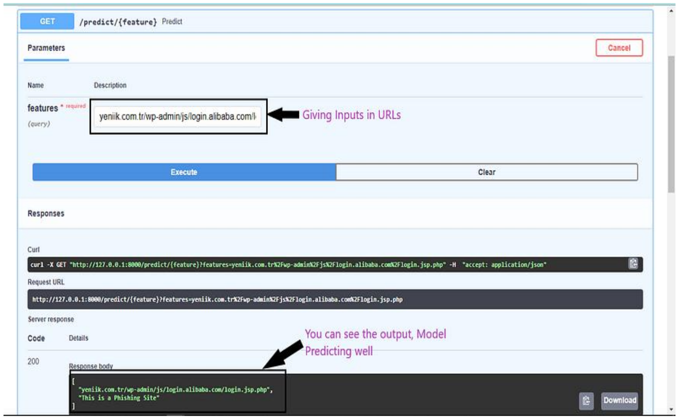
```
http://127.0.0.1:8000/predict/{feature}?features=https%3A%2F%2Fwww.youtube.com%2Fwatch%3Fv%3DqI0TQJI3vdU
```

You can see the Output, Model Predicting well.

200 Response body

```
[
  "https://www.youtube.com/watch?v=qI0TQJI3vdU",
  "This is not a Phishing Site"
]
```

Genuine Site



features * required (query): yeniik.com.tr/wp-admin/js/login.alibaba.com/l

Giving Inputs in URLs

Curl

```
curl -X GET "http://127.0.0.1:8000/predict/{feature}?features=yeniik.com.tr%2Fwp-admin%2Fjs%2Flogin.alibaba.com%2Flogin.jsp.php" -H  "accept: application/json"
```

Request URL

```
http://127.0.0.1:8000/predict/{feature}?features=yeniik.com.tr%2Fwp-admin%2Fjs%2Flogin.alibaba.com%2Flogin.jsp.php
```

You can see the output, Model Predicting well

200 Response body

```
[
  "yeniik.com.tr/wp-admin/js/login.alibaba.com/login.jsp.php",
  "This is a Phishing Site"
]
```

VI. Conclusion & Future Scope

As we already know, Email has become the most important mode of communication in recent years; with internet access, any message can be delivered anywhere in the world. Spam emails, also known as non-self, are unwanted commercial or malicious emails that affect or hack personal information such as bank information, money-related information, or anything that causes destruction to a single individual, a corporation, or a group of people. Aside from advertisements, these may contain links to phishing or malware hosting websites designed to steal sensitive information [6]. Future Scope for this would be to dive deeper into the spam mails and check its attachments and check if the attachment is malicious or not. Spam is a serious problem that is not only irritating to end users, but also financially damaging and a security risk. As a result, this system is designed in such a way that it detects and prevents unsolicited and unwanted emails, thereby aiding in the reduction of spam messages, which would be beneficial to both individuals and the company. In this semester we have trained the machine learning model to successfully identify spam and not spam text used in email body. In the next semester we are planning to improve the model by using more algorithms and databases to improve the accuracy and also try scanning the attachments which are sent along with emails using Natural language processing to further avoid any cyber-attack using emails.

VII. References

[1] R. Abdulraheem, A. Odeh, M. Al Fayoumi and I. Keshta, "Efficient Email phishing detection using Machine learning," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), 2022, pp. 0354-0358, doi: 10.1109/CCWC54503.2022.9720818.

[2] T. M. Ma, K. YAMAMORI and A. Thida, "A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification," 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), 2020, pp. 324-326, doi: 10.1109/GCCE50665.2020.9291921.

[3] N. Kumar, S. Sonowal and Nishant, "Email Spam Detection Using Machine Learning Algorithms," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 108-113, doi: 10.1109/ICIRCA48905.2020.9183098.

[4] L. Shalini, S. S. Manvi, N. C. Gowda and K. N. Manasa, "Detection of Phishing Emails using Machine Learning and Deep Learning," 2022 7th International Conference on Communication and Electronics Systems (ICCES), 2022, pp. 1237-1243, doi: 10.1109/ICCES54183.2022.9835846.

[5] Areej Alhogail, Afrah Alsabih,Applying machine learning and natural language processing to detect phishing email,Computers & Security,Volume 110,2021,102414,ISSN 0167-4048,https://doi.org/10.1016/j.cose.2021.102414.(https://www.sciencedirect.com/science/article/pii/S0167404821002388)

[6] A. Sandra Grace and R. S. Nisha, "Using Python and Machine Learning Algorithms to Create a Spam Classifier," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), 2022, pp. 556-562, doi: 10.1109/CCiCT56684.2022.00103.

[7] M. Altamash and S. N. Singh, "Reconnaissance of Credentials through Phishing Attacks & it's Detection using Machine Learning," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), 2022, pp. 350-358, doi: 10.1109/COM-IT-CON54601.2022.9850698.

[8] ASHUTOSH PRASAD BHATT and Dr.MONIKA SHARMA, " E-mail Security Framework Through Various Virus Encryption Techniques," 2019 International Conference on Intelligent Computing and Control Systems (ICICCS), 2019.

[9] Rattanachai Wongwatkit, Montree Raktham, Thanadet Phawananthaphuti, " Intelligent Blacklist Security System for Protecting Spammer in Corporate Email Solution: A Case of Corporate Email Service Provider in Thailand," International Conference on Advanced Communications Technology (ICACT); 2021.

[10] GEORGIOS KAMBOURAKIS, GERARD DRAPER GIL AND IGNACIO SANCHEZ, " What Email Servers Can Tell to Johnny: An Empirical Study of Provider-to-Provider Email Security," 2019 n, Joint Research Centre (JRC), 2020.

[11] ASIF KARIM, SAMI AZAM, BHARANIDHARAN SHANMUGAM, KRISHNAN KANNOORPATTI AND MAMOUN ALAZAB, " A Comprehensive Survey for Intelligent Spam Email Detection," College of Engineering, IT and Environment, Charles Darwin University, 2019, doi: 10.1109/GCCE50665.2019.2954791.

[12] Kriti Agarwal and Tarun Kumar, " Email Spam Detection using integrated approach of Naïve Bayes and Particle Swarm Optimization," 2018 International Conference on Intelligent Computing and Control Systems (ICICCS), 2018

[13] Akash Iyengar, G.Kalpana, Kalyankumar.S, S.GunaNandhini, "Integrated Spam Detection for Multilingual Emails," INTERNATIONAL CONFERENCE ON INFORMATION, COMMUNICATION & EMBEDDED SYSTEMS (ICICES), 2017

[14] Dhanushka Niroshan, and Tharindu Shehan Ranaweera, "NoFish; Total Anti-Phishing Protection System" *2020 International Conference on Advancements in Computing (ICAC)*, 2020.

[15] Shweta Singh, M.P. Singh, Ramprakash Pandey, "Phishing Detection from URLs Using Deep Learning Approach," *2020 IEEE Global Engineering Education Conference (EDUCON)*, 2020.

[16] YONG FANG , CHENG ZHANG, CHENG HUANG , LIANG LIU, AND YUE YANG, "Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism," College of Cybersecurity, Sichuan University, 2019.

[17] Rabab Alayham Abbas Helmi, and Muhammad Irsyad Abdullah. "Email Anti-Phishing Detection Application." *9th IEEE International Conference on System Engineering and Technology* , 2019

[18] Chirag Bansal and Brahmaleen Sidhu, " Machine Learning based Hybrid Approach for Email Spam Detection," 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021

[19] Dr. A. Sumithra, A. Ashifa, S. Harini, N. Kumaresan, "Probability-based Naïve Bayes Algorithm for Email Spam Classification," International Conference on Computer Communication and Informatics (ICCCI), 2022.

[20] Sanaa Kaddoura, Omar Alfandi, Nadia Dahmani, "A Spam Email Detection Mechanism for English Language Text Emails Using Deep Learning Approach," IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2020.

[21] Nurafifah Alya Farahisya and Fitra A. Bachtiar "Spam Email Detection with Affect Intensities using Recurrent Neural Network Algorithm," 2nd International Conference on Information Technology and Education (ICIT&E), 2022.

[22] Rabab Alayham Abbas Helmi, and Muhammad Irsyad Abdullah. "Email Spam Detection using Deep Learning Approach," International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), 2022

[23] Shubhangi Suryawanshi, Anurag Goswami, Pramod Patil, " Email Spam Detection : An Empirical Comparative Study of Different ML and Ensemble Classifiers," INTERNATIONAL CONFERENCE ON INFORMATION, COMMUNICATION & EMBEDDED SYSTEMS (ICICES), 2019

[24] Dhanushka Niroshan, and Tharindu Shehan Ranaweera, "Email Classification using LSTM: A Deep Learning Technique" *2021 International Conference on Cyber Warfare and in Security (ICCWS)*, 2021.

[25] Jianghong Wei, Xiaofeng Chen, Jianfeng Wang , Xuexian Hu , and Jianfeng Ma, "Enabling (End-to-End) Encrypted Cloud Emails With Practical Forward Secrecy," IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, 2022.