International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 06 | Jun 2023 www.irjet.net p-ISSN: 2395-0072

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

[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.