RESEARCH ARTICLE OPEN ACCESS

Precise Email Classification using Deep Learning

Ms. Deepali Bhimrao Chavan
Computer Science & Engineering Department AMGOI, Vathar- Maharasthra
Prof. Suraj Shivaji Redekar
HOD, Computer Science & Engineering Department AMGOI, Vathar - Maharasthra

ABSTRACT
In practically every industry today, from business to education, emails are used. Ham and spam are the two subcategories of emails. Email spam, often known as junk email or unwelcome email, is a kind of email that can be used to hurt any user by sapping their time and computing resources and stealing important data. Spam email volume is rising quickly day by day. Today's email and IoT service providers facehuge and massive challenges with spam identification and filtration. Email filtering is one of the most important and well-known methods among all the methods created for identifying and preventing spam. SVM, decision trees, CNN, and other machine learning and deep learning approaches have all been applied to this problem.
Together with the explosive growth in internet users, email spam has increased substantially in recent years. Individuals are using them for illegal and dishonest purposes, such as fraud, phishing, and distributing malicious links through unsolicited email that can harm our systems and attempt to access your systems. By quickly constructing phony profiles and email accounts, spammers prey on those who are ignorant of these scams. They use a real name in their spam emails. As a result, it's critical to identify spam emails that include fraud. This project will accomplish this by utilizing machine learning methods, and this article will examine the machine learning algorithms, put them to useon our data sets, and select the approach that can detect email spam with the maximum degreeof precision and accuracy.

I. INTRODUCTION

Email spam, often known as electronic mail spam, is the practice of sending unwanted emails or commercial emails to a list of subscribers. Unsolicited emails signify that the recipient has not given consent to receive them. Throughout last decade, using spam emails has grown in popularity. Spam has grown to be a significant online problem. Spam wastes space, time, and message delivery. Although automatic email filtering may be the best way to stop spam, modern spammers may quickly get around all of these apps. Prior to a few years ago, the majority of spam that came from particular email addresses could be manually stopped.
For spam detection, a machine learning approach will be utilised.

The most popular form of official communication for business purposes is email. Despite the existence of other communication tools, email usage keeps growing. Today's environment necessitates automated email management due to the daily increase in email volume. More than 55% of the emails overall are classified as spam. This demonstrates how these spams waste email users' time and resources while producing nothing helpful. Because spammers utilize sophisticated and inventive tactics to carry out their criminal actions through spam emails, it is crucial to comprehend the various spam email classification approaches and how they work.
Emails are a popular form of communication for both personal and business purposes. An intelligent and successful approach for detecting phishing websites is built on applying classification or association Data Mining methods. The suggested system can be thought as as a classification issue with two categories, ham and phished, with the purpose of detecting phished email. In the area of artificial intelligence known as machine learning, the system is given the capacity to learn without being specifically designed. Algorithms for supervised machine learning are utilized for classification in our model.

1.1 PROBLEM STATEMENT:

Several packet features were extracted for categorization purposes from each email in a self-created dataset that included n phished emails and m ham emails. The classifiers receive these features, and the results are recorded. The goal is to categories data using a variety of machine learning techniques while employing the fewest possible features to create a system that is more accurate.

1.2 OBJECTIVE:

1. Create and implement a machine learning strategy for email phishing detection using huge synthetic as well as real-time data.

2. to create a strategy utilizing different machine learning algorithms and investigate the accuracy using majority routing.

3. create an algorithm to extract various features from emails in order to improve classification accuracy.

4. to examine and verify the system's classification findings using current detection methods.

II. RELATED WORK

Nikhil Kumar, Sanket Sonowal, Nishant et.al [1] Email spam has grown significantly in recent years along with the rapid expansion of internet users. They are being used for fraud, phishing, and other unethical and criminal activities. sending dangerous links through unsolicited emails, which might damage our systems and enter your systems.

Naeem Ahmed, Rashid Amin ,Hamza Aldabbas et.al [2] In practically every industry today, from business to education, emails are used. Ham and spam are the two subcategories of emails. Email spam, often known as junk email or unwelcome email, is a kind of email that can be used to hurt any user by sapping their time and computing resources and stealing important data.

Luo GuangJun, Shah Nazir,Habib Ullah Khan,and Amin Ul Haq et.al [3] The detection of spam is a significant problem in mobile SMS communication, which makes it insecure. A precise and accurate mechanism for detecting spam in mobile SMS communication is required to address this issue. For accurate identification, we suggested using machine learning-based spam detection techniques.

Sridevi Gadde; A. Lakshmanarao; S. Satyanarayana[4]Those who use mobile devices are becoming more numerous every day. Both smartphones and basic phones support SMS (short message service), a text messaging service. As a result, SMS traffic dramatically rose. There were also more spam messages. The spammers attempt to send spam communications in order to benefit financially or commercially, such as market expansion, collection of credit card information, etc. Hence, spam classification is given considerable consideration. In this study, we used a variety of machine learning and deep learning techniques to identify SMS spam. We created a spam detection model using data from UCI.

Mehul Gupta ; Aditya Bakliwal; Shubhangi Agarwal; Pulkit Mehndiratta [5]Short Messaging Service (SMS) usage on phones has expanded to such a big degree due to technological developments and an increase in content-based advertising that devices are occasionally inundated with a large number of spam SMS. Private data loss is another risk posed by these spam mailings. There are numerous content-based machine learning methods that have been successfully used to filter spam emails. Contemporary studies have classified text messages as spam or gammon using certain stylistic characteristics. The ability to detect SMS spam can be significantly impacted by the use of well-known terms, phrases, abbreviations, and idioms.

Asma Bibi1, Rasia Latif1, Samina Khalid1*, Waqas Ahmed2, Raja Ahtsham Shabir1, Tehmina Shahryar et. Al [6] Emails are a common form of communication on both a personal and business level. As time goes on, emails are increasingly being used for spamming, distributing viruses, and defrauding internetusers.

Some acceptable emails are classified as ham, whereas certain kinds of unsolicited emails are classified as spam.

Many machine learning techniques are utilised during the course of the year to estimate the category of emails. In this essay, we consider a classifier that is effective at classifying texts.

Neelam Choudhary,Ankit Kumar Jain et.al [7] Mobile devices are becoming more and more popular since they offer a wide range of services at lower prices. SMS, or short message service, is one of the more popular types of communication. However, this has increased attacks on mobile devices, such as SMS spam. In this research, we describe a novel strategy that makes use of machine learning classification techniques to identify and filter spam communications. We have examined the characteristics of spam messages before identifying ten features that can effectively separate SMS spam from ham transmissions.

EmmanuelGbenga Dada, Joseph Stephen Bassi, Haruna Chiroma , Shafi'i Muhammad Abdulhamid, Adebayo Olusola Adetunmbi, Opeyemi Emmanuel Ajibuwa Et.al [8] The need for more dependable and powerful antispam filters has increased dramatically due to the rise in the number of unsolicited emails, or spam. Recently, spam emails have been successfully detected and filtered using machine learning techniques. We give a thorough analysis of a few well-liked machine learning-based email spam filtering techniques. Our analysis includes a summary of the key ideas, initiatives, successes, and current research directions in spam filtering. The study background's preliminary discussion looks at how machine learning techniques are applied to the email spam filtering systems of the top

internet service providers (ISPs), including Google, Yahoo, and Outlook.

Paras Sethi; Vaibhav Bhandari; Bhavna Kohli et.al [9] Spam emails and messages have increased during the last few years. Today, spam text messages can be combated with the help of legal, economic, and technical methods. Bayesian filters play a significant part in preventing this issue. In order to identify spam messages delivered on mobile devices, we examined and compared the relative merits of various machine learning techniques in this research. For our testing and validation needs, we built two datasets using data from an available public dataset.

S. Nandhini; Jeen Marseline K.S. et.al [10] Sending a great deal of unwanted email puts consumers' security at risk. Despite numerous security measures, spammers significantly increase internet vulnerability. The effective use of various well-known algorithms for creating a machine learning model that can distinguish between spam and legitimate mail is covered in this work. UCI The experiment uses the Machine Learning Repository Spambase Data Set. In order to train and develop a powerful machine learning model for email spam detection, the performance of five significant machine learning classification algorithms, including Logistic Regression, Decision Tree, Naive Bayes, KNN, and SVM, is assessed. The data set is trained and tested using the Weka tool.

III. PROPOSED SYSTEM:

Email analysis is typically one of its most frequent actions and is categorized under text analysis.
Algorithms used in email analysis include those like CNN, SVM, and decision trees. Classifying emails as spam rather than non-spam is a key analysis topic in email categorization.
Various publications on email spam classification and analysis attempted to categories emails. Spam supported the sender's gender given a lot of the characteristics that make emails from women or men distinct from one another.
Emails can be divided into two categories relative to spam and non-spam emails: attention-grabbing emails and dull emails.
Email grouping also included the idea of grouping emails into different folders or subjects.
Also, several research studies use the temporal information contained in emails (such as when they were sent, received, etc.) to analyse emails.
Several analysis papers make an effort to explain why emails supported similar topics or subjects.
Certain email systems, like Google, group emails that are connected to one another.

IV. ALGORITHM

Support Vector Machine (SVM):
One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary. SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method.
SVM comes in two varieties:
1. Linear SVM: Linear SVM is used for data that can be divided into two classes using a single straight line. This type of data is called linearly separable data, and the classifier employed is known as a Linear SVM classifier.

2. Non-linear SVM: Non-Linear SVM is used for non-linearly separated data. If a dataset cannot be classified using a straight line, it is considered non-linear data, and the classifier employed is referred to as a Non-linear SVM classifier.

Convolutional Neural Network (CNN):
Convolutional Neural Networks are designed specifically for use in image and video recognition applications. CNN is primarily utilized for image analysis applications such segmentation, object detection, and picture recognition.
Convolutional Neural Networks have four different kinds of layers:
1) Convolutional Layer: Each input neuron in a conventional neural network is connected to the following hidden layer. Only a small portion of the input layer neurons in CNN are connected to the hidden layer of neurons.
2) Pooling Layer: The pooling layer is used to make the feature map less dimensional. Inside the CNN's hidden layer, there will be numerous activation and pooling layers.
3) Flatten: Flattening is the process of reducing data to a 1-dimensional array so that it may be entered into the following layer. We flatten the

convolutional layer output to produce a solitary, lengthy feature vector.

4) Fully Connected Layer: Fully Connected Layers make up the network's final few tiers. The output from the last pooling or convolutional layer is passed into the fully connected layer, where it is flattened before being applied.

Decision Tree:
A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favored for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. The Decision Node and Leaf Node are the two nodes of a decision tree. Whereas Leaf nodes are the results of decisions and do not have any more branches, Decision nodes are used to create decisions and have numerous branches.

The given dataset's features are used to execute the test or make the decisions. It is a graphical depiction for obtaining all feasible answers to a choice or problem based on predetermined conditions. It is known as a decision tree because, like a tree, it begins with the root node and grows on subsequent branches to form a structure resembling a tree. The CART algorithm, which stands for Classification and Regression Tree algorithm, is used to construct a tree. A decision tree only poses a question and divides the tree into subtrees according to the response (Yes/No).

V. ADVANTAGES

1. Extraction of link-based features from emails.
2. For the complete dataset, tag-based features extraction
3. Extracting word base characteristics.
4. All test data were classified as either fishing or normal, accordingly.

VI. CONCLUSION

spam detection is crucial for protecting email and message communication. A significant problem is the accurate identification of spam, and numerous detection techniques have been put forth by various researchers. Nevertheless, these techniques fall short in their ability to correctly and effectively detect spam. We have suggested a technique for spam identification using machine learning predictive models to address this problem. Consequently, the findings imply that the suggested approach is more trustworthy for precise and prompt identification of spam and will secure messaging and email systems.

REFERENCE

1. “Email Spam Detection Using Machine Learning Algorithms ”https://ieeexplore.ieee.org/
2. “Machine Learning Techniques for Spam Detection in Email and IoT Platforms” https://www.hindawi.com/
3. “Spam Detection Approach for Secure Mobile Message Communication Using Machine Learning Algorithms” https://www.hindawi.com/
4. “SMS Spam Detection using Machine Learning and Deep Learning Techniques” https://ieeexplore.ieee.org/
5. “A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers” https://ieeexplore.ieee.org/
6. “Spam Mail Scanning Using Machine Learning Algorithm” http://www.jcomputers.us
7. “Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique ”https: //link.springer.com
8. “Machine learning for email spam filtering: review, approaches and open research problems ”https
9. ”SMS spam detection and comparison of various machine learning algorithms” https://ieeexplore.ieee.org
10. ”Performance Evaluation of Machine Learning Algorithms for Email Spam Detection”
11. K. Krombholz, H. Hobel, M. Huber, and E. Weippl “Advanced Social Engineering Attacks”, Journal of information security and applications 22 (2015) 113-122
12. U. H. Rao and U. Nayak, “The InfoSec Handbook”, Chapter 15 Social Engineering, 01 September 2014 pages 307-323
13. T. Ayodele, C. Shoniregun, and G. Akmayeva, “Anti-Phishing Prevention Measure for Email Systems”, World Congress on Internet Security (WorldCIS-2012)
14. C. Olivo, A. Santin, and L. Oliveira, “Obtaining the threat model for email phishing”, Applied Soft Computing 13 (2013) 4841–484