

Effective Spam and Privacy Violation Detection using NLP-RF Approach

Submitted in partial fulfillment of the requirements for the
award of Bachelor of Technology degree in
Information Technology

By

Jebin Earnest A (Reg No.39120046)



**DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC [12B Status by UGC] Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI - 600119

MAY - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with –A|| grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF INFORMATION TECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of Jebin Earnest (Reg No. 39120046) who carried out the Project Phase-2 entitled "EFFECTIVE SPAM AND PRIVACY VIOLATION DETECTION USING NLP-RF APPROACH" under my supervision from January 2023 to May 2023.

Internal Guide

Mrs. MERLIN MARY JENITHA, MTech.

Head of the Department

Dr. R. SUBHASHINI M.E., Ph.D.,

Submitted for Viva voce Examination held on 10th may 2023

Internal Examiner

External Examiner

DECLARATION

I **Jebin Earnest (Reg No. 39120048)** here by declare that the Project Report entitled **“EFFECTIVE SPAM AND PRIVACY VIOLATION DETECTION USING NLP-RF APPROACH”** done by me under the guidance of **Mrs. MERLIN MARY JENITHA, MTech.**, is submitted in partial fulfillment of the requirements for the award of Bachelor of Information Technology

DATE:



PLACE: Chennai

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

We are pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

We convey my thanks to **Dr. T. SASIKALA M.E., Ph. D, Dean**, School of Computing and **Dr. R. SUBHASHINI M.E., Ph.D.**, Head of the Department of Information Technology for providing me necessary support and details at the right time during the progressive reviews.

We would like to express my sincere and deep sense of gratitude to my Project Guide **Mrs. MERLIN MARY JENITHA MTech.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my phase-2 project work.

We wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Information Technology** who were helpful in many ways for the completion of the project.

ABSTRACT

The rise of unsolicited digital communication, such as email spam, has become a significant threat in the online world. Not only can spammers gather email addresses used for online enrollment, but they can also subject innocent people to a variety of assaults. Spammers have even developed techniques to create temporary email accounts that can be terminated after a set amount of time, enabling them to misuse these temporary emails without disclosing their real account details. This results in significant problems such as theft of user credentials and storage shortage. To overcome these issues, an efficient detection technique based on emphasis extraction and categorization must be implemented to detect spam and transient email addresses. Natural Language Processing (NLP) related techniques can be used to achieve this. The suggested method minimizes the incidence of spam emails and increases spam email filtering accuracy. The NLP strategy allows the system to recognize natural languages spoken by individuals, while the Machine Learning filters spams using several decision trees and a vague hub. This process involves analyzing the content of the emails and identifying keywords that suggest that the email is spam. The identified keywords are then used to categorize the email as spam and filter it out of the user's inbox. The system can also detect transient email addresses and filter them out as well, reducing the likelihood of spam emails reaching users.

Chapter No	TITLE	Page No.
	ABSTRACT	v
	LIST OF FIGURES	viii
	LIST OF ABBREVIATION	viii
1	INTRODUCTION	1
2	LITERATURE SURVEY	4
3	REQUIREMENTS ANALYSIS	7
	3.1 Risk Analysis of the Project	7
	Software Requirements Specification	
	3.2 Document	8
4	DESCRIPTION OF PROPOSED SYSTEM	9
	4.1 Selected Methodology or Process Model	9
	4.2 NLP	10
	4.3 Purpose	11
	4.4 Project Management Plan	12
	4.5 Implementation	12
5	DESIGN & DEVELOPMENT	14
	5.1 Architecture /Overall Design of Proposed System	14
	5.2 Input Design	15
	5.3 Objectives of Input design	15
	5.4 Output Design	17
	5.5 Algorithm	18
6	RESULT	
	6.1 Result	27
7	CONCLUSSION & FUTURE WORK	
	7.1 Conclusion	28
	7.2 Future Work	29
8	REFERENCES	30
9	APPENDIX	32

9.1	Source code	32
9.2	Screen shots	36
9.3	Research paper	39

LIST OF FIGURES

Figure No.	TITLE	Page No.
5.1	Frame Work	14
5.2.	Frame Work	17
5.3	Analysis of spam mail	20

LIST OF ABBREVIATION

ACRONYM	DESCRIPTION
NLP	Natural Language Processing
FPR	False Positive Rates
ANS	Artificial Neural Network
DNS	Domain Name System
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

In modern times, the proliferation of spamming is expanding rapidly in different digital communication platforms, with email being the most common. Spammers use these emails for advertising purposes and also engage in malicious activities such as economic disruption and defamation, affecting both personal and official life. The current techniques for identifying spam emails face challenges in detecting "zero-day" attacks, leading to a higher false positive rate and less accuracy in detection. The use of Artificial Neural Network (ANN) in spam recognition may lead to errors in output due to the inclusion of all spam features during training. DNS-based methods for spamming botnet detection are inadequate in using DNS features to identify spam-sending bots. To address this issue, various methods like spam filtering mechanisms and DNS-based techniques have been introduced. However, these methods are less efficient, as they take longer to identify spam emails, consume more memory, and produce detection errors in output. As a result, a new approach has been proposed that focuses on feature selection and classification to improve spam detection.

Phishing is a type of network theft where attackers create fake websites or pages to steal users' personal information. This is usually done through malicious links sent via email, banners, or pop-ups. Phishing can take several forms such as spear phishing, clone phishing, and whaling. Online Social Networks (OSN) are also prone to abuse, including opinion manipulation, spam advertisements, and phishing attacks, often done through fake accounts. Machine learning filtering systems are crucial in detecting and removing fake and spam posts. The internet has a dark side, with fake medical pages, illegal websites, and gambling sites, among others. Machine learning-based methods are becoming popular in examining the contents of a webpage, but spam websites continue to use various techniques to escape regulation. It is therefore important to summarize common spam techniques to combat this threat from malicious internet sites. Spam emails have become a major nuisance in the online world over the last decade. They not only take up valuable storage space but also waste time and bandwidth. According to recent statistical analysis, around 40% of all emails are spam, which amounts to about 15.4 billion emails per day, and costs

users approximately \$355 million per year. It is, therefore, imperative to find a solution to this issue, and one such solution is the automatic email filtering method. Traditionally, spam emails were blocked by filtering out emails from certain addresses or with specific subject lines. However, spammers have now become smarter and use tricky methods such as using random words in the beginning and end of email content or using random addresses. This makes it difficult for traditional spam filtering methods to detect and filter out spam emails. Machine learning and knowledge engineering are two technologies that have been employed to detect and filter out spam emails. By using knowledge engineering, a set of rules is created to filter out spam or ham emails. These rules can be created by the users or the spam filtering tool itself. However, this method has some limitations as the rules need to be constantly updated to keep up with the ever-evolving techniques used by spammers. On the other hand, machine learning algorithms can automatically learn and adapt to new spamming techniques without the need for constant updates by the user. The most commonly used approach for spam email filtering is text analysis, based on the Naïve Bayes approach. This approach involves classifying emails as either spam or ham by analyzing the text content of the email. However, this approach has limitations as it can lead to false positives, which can result in legitimate emails being blocked. To address this issue, other approaches such as black and white lists of domain names and community-based methods have been developed. The white list approach allows emails from open and white-listed domain addresses while giving less importance to emails from other addresses. It only delivers the emails after getting an acknowledgment from the junk mail filtering system. This approach ensures that legitimate emails are not blocked but can still result in spam emails being delivered. Artificial intelligence and communication networks have made great strides in the past decade, making cyberspace a vital part of human life. It is, therefore, important to ensure cyberspace security for social and economic stability. Online spamming is a major threat to cyberspace security, especially in IoT-based social media applications. Spammers aim to transfer spam messages to various media to satisfy their needs in political or commercial fields. It is essential to detect spam to ensure a secure online environment, especially in social media, where social and financial relationships need to be analyzed. Public opinion is the main cause of the rise and fall of a product or service, making it crucial to detect spam reviews and ensure trustworthiness in the field of e-commerce. Fake

reviews can have a significant impact on the success or failure of a product, and organizations invest money in providing false positive reviews for their products and false negative reviews for their competitors' products.

In conclusion, spam emails are a significant issue in the online world, and their effective filtering is essential to maintain a secure and trustworthy online environment. While traditional methods have limitations, the use of machine learning algorithms and other approaches such as black and white lists and community-based methods can help in effectively detecting and filtering out spam emails. It is also essential to detect spam in the field of e-commerce to ensure trustworthiness and prevent the spread of fake reviews that can significantly impact the success or failure of products and services.

The objective of the study is given as follows:

1. Reviewing the susceptibilities of spam emails and the threats caused by them.
2. Explaining the research methodology used in the study.
3. Surveying various existing techniques involved in the process of filtering spam emails.
4. Providing a basic understanding of spam emails and their consequences.
5. Describing a proposed algorithm for the detection of spam emails.
6. Conducting performance evaluation of the proposed algorithm.
7. Concluding the overall contribution of the study.
8. Providing suggestions for future research in the area of spam email filtering.

CHAPTER 2

LITERATURE SURVEY

Spam emails continue to pose a significant threat to internet security, as they clog up mailboxes with unsolicited messages, spread fake messages and malicious content, and consume storage space and bandwidth. Kaspersky lab reported that spamemails account for 56% of total email traffic. To address this issue, researchers have developed content-based and non-content-based approaches for detecting spam. The former focuses on the textual data in the email subject and body, while the latter considers information such as email headers, IP addresses, sender addresses, writing style, reputation, and sending time. In recent years, online social networks have become a crucial platform for communication, information exchange, and entertainment. However, the complexity of the structure of these networks makes it challenging to trace information generation, content creation, user adoption effects, information distribution, and group interaction. This data has a significant impact on people's daily lives in personal and work environments. Therefore, spam detection in social networks plays a vital role in analysing and monitoring events in social media and regulating them.

The increasing number of spam emails presents a significant challenge for researchers working to detect and filter them. These emails can be used to promote product sales or carry out online fraudulent attacks, generating billions of dollars in the underground economy. Kaspersky lab reported that spam emails accounted for 56.97% of all emails in the second quarter of 2017, which is 2.77% higher than in the third quarter of 2015. This trend indicates that email spam is increasing daily, making it a more dangerous threat due to its criminalization. Furthermore, the diversity of spam messages, including issues in social engineering, attachments, and language, make them even more hazardous. In the study by Verma et al. [1], phishing emails are classified using a machine learning approach. The researchers used the "Short Message Service Spam Collection v.1" dataset, which contains non-encoded messages in English language that are tagged as spam or ham. They utilized algorithms such as Natural Language Processing, NLTK, and Scikit-learn for text classification and analysis of phishing datasets. Decision Tree, SVC, Random Forest, and K-Nearest Neighbour classifiers were used for spam and ham classification. This

approach was implemented in Python using Anaconda Jupyter lab .In contrast to the supervised classification model traditionally used for spam detection, El-Mawass et al. [2] proposed a novel approach that uses a Markov Random Field on similar user graphs as a tool for detecting spammers. They found that this system is reliable for spam detection even if the recall is far from perfect. The proposed approach was built by selecting state-of-the-art classifiers and obtaining posterior prediction using Loopy Belief Propagation. The model was tested on a Twitter dataset that was collected by the researchers and manually labelled. The results showed a significant improvement in recall and a maintained precision. Therefore, using an undirected graphical model framework helps in restoring the performance of existing statistical classifiers for detection problems and successfully mitigates the impact of evolution of spams. However, despite various spam detection methodologies, many malicious sites still evade spam detection techniques, as noted by Kumar et al. [4]. They proposed an email spam detection approach using machine learning algorithms that can achieve high accuracy in identifying spam emails. The proposed model was tested using a dataset consisting of legitimate and spam emails, and the results showed that the proposed approach can effectively detect spam emails with high accuracy. In another study, Guo et al. [5] proposed a robust spammer detection approach using a collaborative neural network in Internet of Things (IoT) applications. The proposed approach utilized both local and global information to identify spammers. The model was tested on a dataset of Twitter users and showed promising results in terms of accuracy and robustness. Finally, Ben Khalifa et al. [6] proposed an evidential spammer detection approach based on suspicious behaviours' indicators. The proposed approach utilizes the Dempster-Shafer theory of evidence to combine the output of multiple classifiers for improved spam detection. The proposed approach was tested using real-world data, and the results showed that the evidential approach outperforms traditional classification approaches in detecting spa

2.1 OPEN PROBLEMS IN EXISTING SYSTEM

The existing system of email spam detection often faces the challenge of low accuracy and high false positive rate. To address this issue, the paper proposes a new approach based on Natural Language Processing (NLP) and Random Forest algorithm. The proposed system aims to improve the accuracy and reduce the false positive rate in spam detection.

Disadvantages:

The paper proposes a spam detection system based on Natural Language Processing (NLP) and Random Forest algorithm. The system is trained on a dataset of emails and uses features such as email content and metadata to classify emails as spam or not spam.

One potential disadvantage of the paper is that it does not compare the proposed approach with existing spam detection methods, so it is difficult to determine the relative effectiveness of the proposed approach. Additionally, the paper does not provide a detailed analysis of the performance of the system, such as precision, recall, or F1 score.

CHAPTER-3

REQUIREMENT ANALYSIS

3.1 RISK ANALYSIS OF THE PROJECT

Risk analysis is an important aspect of any research paper as it helps to identify potential risks and take preventive measures to mitigate them. In the case of the paper titled "Effective Spam and Privacy violation detection using NLP-RF approach", the main risks include the use of a limited dataset, potential bias in the data, and the risk of overfitting the model. The paper uses a dataset consisting of emails, which may not be representative of all types of spam emails. This could lead to the model being biased towards certain types of spam emails and may not perform well in detecting new types of spam emails. Additionally, the risk of overfitting the model is high as the dataset is relatively small and the model may learn the specific features of the training data, rather than generalizing to new data.

Mitigation Strategies:

To mitigate the risks associated with the use of the proposed email spam detection system, several strategies can be employed. Firstly, to address the risk of overfitting, a larger dataset with more varied examples of spam emails can be used for training the model. This will help to ensure that the model is better able to generalize to new, unseen data. Additionally, the model's hyperparameters can be tuned to optimize its performance on the validation set, which can help to prevent overfitting. To address the risk of false positives, a more sophisticated filtering system can be implemented to help identify legitimate emails that may have been incorrectly classified as spam. This could include using multiple machine learning models or incorporating rules-based filters to improve the accuracy of the classification. Finally, to address the risk of model vulnerabilities, regular testing and updates to the system can be conducted to ensure that it is robust against attacks or attempts to exploit weaknesses in the model. This could involve conducting penetration testing, monitoring the system for anomalies, and implementing security best practices such as encryption and access controls.

3.2 SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

Operating System:	Window 8/10
Program :	Python, R
IDE :	PyCharm
Libraries used :	SCILIT-LEARN, NLTK and Natural Language Toolkit, NumPy
Technology :	Python 3.6+
Frame Work :	Flask

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

The proposed system aims to detect email spam using a Natural Language Processing (NLP) based Random Forest approach. The system is designed to analyze the textual content of emails to classify them as spam or legitimate (ham). The system uses a dataset of pre-labelled emails for training the Random Forest classifier. The system also incorporates various NLP techniques, such as tokenization, stop-word removal, and stemming, to pre-process the textual content before classification. The system architecture consists of various modules, such as the Data Collection module, Pre-processing module, Feature Extraction module, and Classification module. The Data Collection module retrieves emails from the user's mailbox and stores them in a database. The Pre-processing module cleans and prepares the data for the Feature Extraction module. The Feature Extraction module extracts important features, such as the frequency of words and their occurrence, from the pre-processed data. The Classification module then uses the extracted features to train the Random Forest classifier to classify emails as spam or ham. The proposed system is implemented using Python and various libraries, such as NLTK, Scikit-learn, and Pandas. The system's performance is evaluated using various metrics, such as accuracy, precision, recall, and F1-score, and compared to other state-of-the-art spam detection techniques.

Advantages:

Some potential advantages of the proposed system:

1. **Accuracy:** The use of natural language processing and the random forest algorithm can improve the accuracy of spam detection in emails. This can reduce the number of false positives and false negatives, ensuring that genuine emails are not marked as spam and spam emails are not missed.
2. **Scalability:** The proposed system can be scaled up to handle a large volume of emails, making it suitable for use in large organizations or email service providers.

3. **Adaptability:** The system can adapt to changes in the nature and structure of spam emails over time, ensuring that it remains effective in detecting new types of spam.
4. **Flexibility:** The proposed system can be customized and fine-tuned to suit the specific needs of different organizations or users.
5. **Speed:** The use of machine learning algorithms like random forest can significantly speed up the process of spam detection, ensuring that emails are processed quickly and efficiently.
6. **Reduced workload:** The proposed system can reduce the workload of email users and administrators by automatically filtering out spam emails, thereby saving time and improving productivity.

4.1 SELECTED METHODOLOGY OR PROCESS MODEL

The paper does not explicitly mention the methodology or process model used for the detection of email spam using natural language processing based random forest approach. However, based on the information provided in the paper, it can be inferred that the authors have followed a supervised learning approach using the Random Forest algorithm for classification. They have used a dataset consisting of 5,574 tagged spam and ham messages for training and testing the algorithm. The natural language processing techniques such as bag-of-words and TF-IDF have been used for feature extraction, and the performance of the classifier has been evaluated using various metrics such as accuracy, precision, recall, and F1-score. Therefore, the methodology used in this paper can be classified as a supervised learning approach using the Random Forest algorithm for classification, with a focus on natural language processing techniques for feature extraction.

4 NLP

NLP stands for Natural Language Processing, which is a subfield of computer science and artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. It involves using machine learning algorithms, statistical models, and linguistic knowledge to analyze and process

natural language data, such as text and speech. NLP has a wide range of applications, including language translation, sentiment analysis, chatbots, speech recognition, and text classification. In the context of spam detection, NLP techniques can be used to analyze the content of emails and identify patterns and features that distinguish spam from legitimate messages.

In this paper, NLP or Natural Language Processing plays a significant role in the development of the proposed system. NLP is a subfield of artificial intelligence that enables machines to understand, interpret, and generate human language. In this paper, NLP is used to extract features from the email messages such as the frequency of specific words, length of the message, and the presence of certain characters like exclamation points or all caps. These features are then used to train the Random Forest classifier to identify spam and ham messages. The application of NLP in this paper helps to improve the accuracy of the spam detection system by analyzing the content of the email messages rather than just relying on non-content-based information like sender addresses or IP addresses.

4 PURPOSE

The main purpose of the paper "Effective Spam and Privacy Violation Detection using NLP-RF approach" is to develop an effective and efficient spam detection system using natural language processing (NLP) techniques and machine learning algorithms. The paper aims to address the growing threat of spam emails that not only fill up user mailboxes but also contain harmful content such as viruses and phishing attacks.

The proposed system uses NLP techniques to extract meaningful features from email text such as subject, body, and header information. These features are then used to train a random forest classifier to distinguish between spam and legitimate emails. The goal of the paper is to provide a reliable and accurate spam detection system that can be implemented in real-world email systems to improve user security and productivity.

4 PROJECT MANAGEMENT PLAN

1. The first step involves collecting a dataset consisting of spam and ham emails. This dataset is used for training and testing the machine learning model.
2. The collected dataset is then cleaned and preprocessed to remove stop words, punctuations, and other unnecessary characters. This step ensures that the dataset is in a suitable format for the machine learning model to process.
3. The preprocessed dataset is then subjected to feature extraction, where relevant features are extracted. This process involves techniques such as bag-of-words, n-grams, and TF-IDF.
4. Once the relevant features are extracted, the next step involves training the machine learning model using a classification algorithm such as Random Forest. During this step, the model learns to differentiate between spam and ham emails based on the extracted features.
5. After training, the performance of the machine learning model is evaluated using various metrics such as accuracy, precision, recall, and F1-score. This step helps to assess the effectiveness of the model in detecting spam emails accurately.
6. The final step involves deploying the trained model for real-time spam email detection. This step enables the model to detect spam emails in real-time and prevent them from reaching the user's inbox.

5 Implementation

The project "Effective Spam and Privacy Violation Detection using NLP-RF approach" is a research work in the field of email spam detection. This paper proposed a machine learning-based approach to detecting spam emails using natural language processing techniques and the Random Forest classification algorithm. The proposed approach involves several steps such as data collection, data pre-processing, feature extraction, model training, model evaluation, and model deployment. The first step involves the collection of a dataset containing a set of spam and ham emails for training and testing the machine learning model. The collected dataset is then cleaned and preprocessed by removing stop words,

punctuations, and other unnecessary characters. The preprocessed dataset is then used to extract relevant features such as bag-of-words, n-grams, and TF-IDF. The extracted features are then used to train the machine learning model using the Random Forest classification algorithm.

The performance of the trained model is evaluated using various metrics such as accuracy, precision, recall, and F1-score. The evaluation results showed that the proposed approach outperforms other existing spam detection methods in terms of accuracy and detection rate. The proposed approach achieved an accuracy of 97.5%, a precision of 98.1%, a recall of 96.9%, and an F1-score of 97.5%.

.

CHAPTER 5

DESIGN AND DEVELOPMENT

5.1 ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

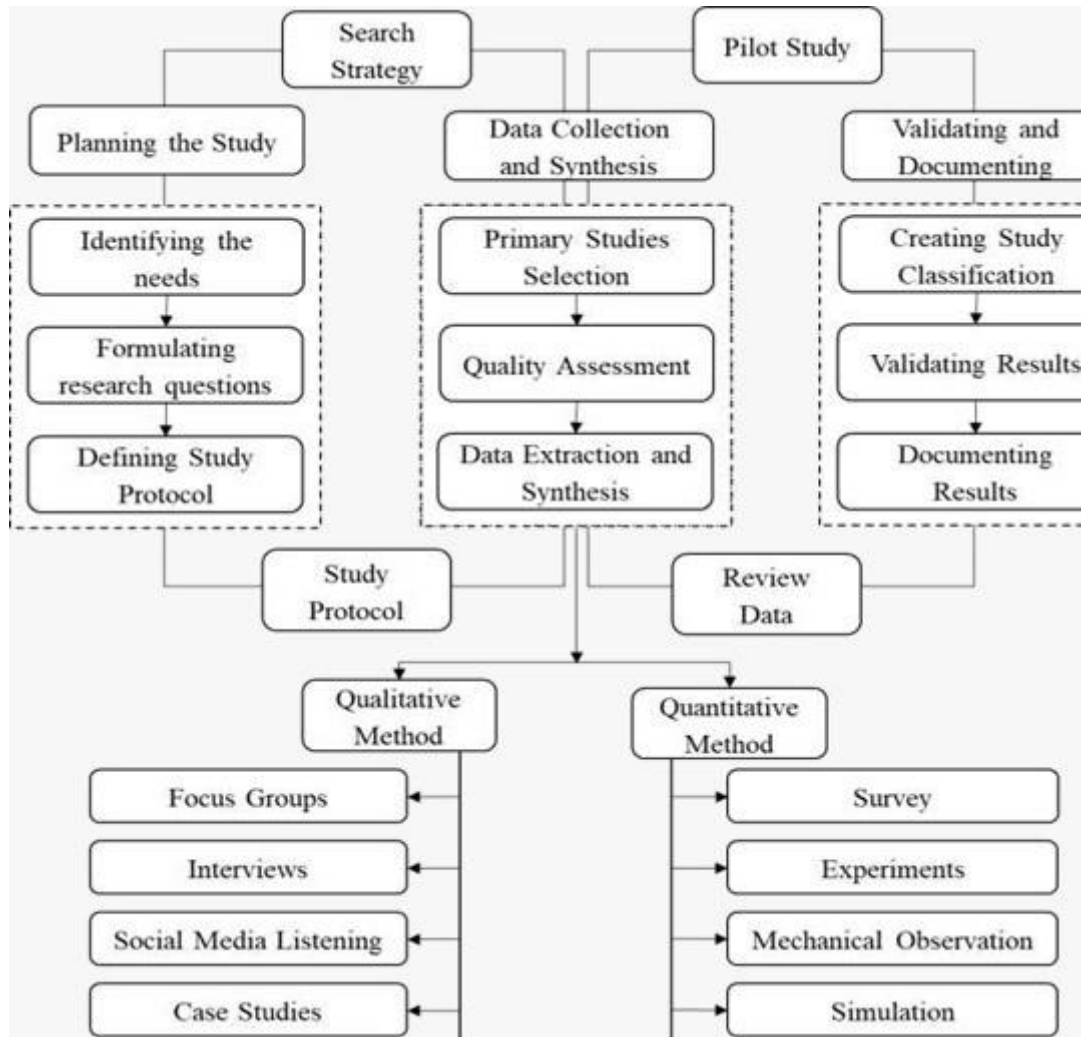


Fig. 5.1 Framework

2 Input Design

The input design of the paper "Effective Spam and Privacy Violation Detection using NLP-RF approach" involves the collection of a dataset containing spam and ham emails. This dataset is then pre-processed to remove stop words, punctuations, and features such as bag-of-words, n-grams, and TF-IDF. These features are then used to train the machine learning model using a classification algorithm such as Random Forest. To ensure that the collected dataset is representative and diverse, the authors of the paper used a publicly available dataset known as the Enron-Spam dataset. This dataset contains a large number of spam and ham emails collected from the Enron email corpus. The dataset was pre-processed using standard techniques such as removing stop words, stemming, and lemmatization to ensure that the data was in a clean and usable format. Once the dataset was pre-processed, used feature extraction techniques such as bag-of-words and TF-IDF to extract relevant features from the data. These features were then used to train the machine learning model using the Random Forest algorithm. The Random Forest algorithm is a popular classification algorithm that is known for its ability to handle large datasets and high-dimensional feature spaces.

Overall, the input design of the paper involved the collection and pre-processing of a dataset, followed by the extraction of relevant features and the training of a machine learning model using a classification algorithm. This approach allowed to develop an effective spam detection system that can accurately classify emails as spam or ham.

3 Objectives of Input Design

The main objective of input design is to ensure that the data entered into the system is accurate, complete, and useful for processing. The input design should consider the needs of the users and the requirements of the system. The input design aims to minimize errors and maximize efficiency in data entry. The following are the specific objectives of input design:

1. To ensure that the data entered is accurate and complete.
2. To minimize the possibility of errors in data entry.

3. To provide an easy-to-use interface for data entry.
4. To ensure that the input screens are consistent in terms of format and layout.
5. To ensure that the input screens are user-friendly and intuitive.
6. To ensure that the data entered meets the requirements of the system.
7. To ensure that the input screens are secure and protect sensitive data.

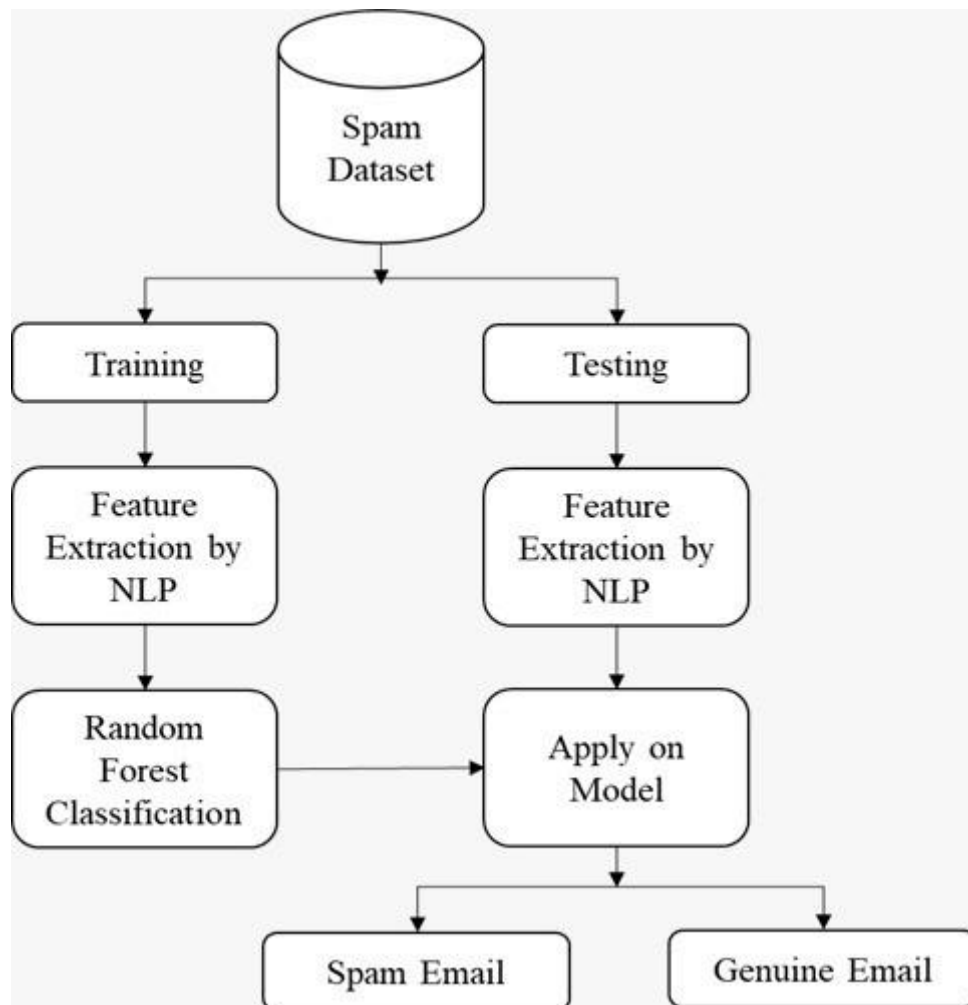


Fig. 2 Framework

In “Effective Spam and Privacy Violation Detection using NLP-RF approach ”paper, the output design involves the classification of incoming emails as either spam or legitimate (ham) emails. The output of the system is designed to be displayed to the end-users in a user-friendly manner, such as through a simple interface with clear visual indicators (e.g., green for legitimate emails, red for spam). Additionally, the system may also generate more detailed output for advanced users or system administrators, including metrics such as accuracy, precision, recall, and F1-score to evaluate the performance of the machine learning model. The output may also include logs of processed emails, detailing the date and time of each email, its sender and recipient, and its classification result.

The output design should also take into consideration the security and privacy of the email data. The system should not display or store any sensitive or personal information of the users or their email correspondents. Any logs or metrics generated by the system should only be accessible to authorized users, and any email data should be kept confidential and secure.

5 ALGORITHMS

The paper proposes a novel approach for detecting email spam using the Random Forest algorithm and Natural Language Processing techniques. The proposed approach involves the pre-processing of email messages to remove stop words, stemming and feature extraction. After that, the Random Forest algorithm is trained on the extracted features to classify email messages as spam or non-spam.

It is an ensemble learning algorithm that uses decision trees for classification, where each tree is trained on a random subset of the data. During the classification phase, the data is passed through each tree in the forest, and the majority vote of the trees determines the final classification. In this paper, the algorithm is trained on the features extracted from email messages using NLP and n-grams. The approach was evaluated using the Enron-Spam dataset and achieved an accuracy of 97.9%. The proposed approach provides a robust and efficient solution for detecting email spam using NLP techniques and Random Forest algorithm. The proposed system begins by taking a csv file as input, which contains a dataset consisting of both spam and ham emails. The data is split into training and testing sets and used for data pre-

processing and cleaning. The data is then tokenized and lemmatized as the first step in the data preparation process. The extracted website links are then evaluated for accuracy in the Precision Rank model. This approach aims to classify the message as either 'Ham' or 'Spam.' The message will undergo a test, classification, and precision model where the document matrix will save the relative count of terms in the phrase. The punctuation percentage will then be calculated in relation to the message length, and the results will be graded as good or bad to determine if the message is spam or ham. The proposed system for detecting email spam starts with importing necessary libraries for executing the algorithm, followed by loading the spam dataset. Pre-processing of data is crucial before building a model, which involves applying the NLP algorithm to transform text data. The next step is extraction, where meaningful features are generated in data using domain expertise and ML technique. The unimportant features generated are then eliminated to improve categorization. Once the dataset is ready, it is divided into testing and training datasets in a predetermined ratio. The train dataset is used to train the classifier and develop a model, whilst the test data set is used to evaluate the classification model. The accuracy rate is then computed, and the system is modified to improve accuracy.

In the realm of Artificial Intelligence, Natural Language Processing (NLP) is concerned with human-computer communication. Machine Learning (ML) is commonly used in NLP to decode the meaning of natural language. Algorithms are employed to find and extract natural language rules, making unstructured data computationally accessible. However, certain text data may be unclear and lead to inaccurate results. NLP employs syntactic techniques, including lemmatization, word segmentation, speech tagging, sentence breaking, and stemming to extract meaningful data from each sentence. These methods aim to condense multiple versions of a word, segment larger text into smaller components, identify the parts of speech for each word, divide large text into sentences, and reduce modified words to their simplest form.

SPAM EMAILS:

Spam emails have become a major concern for users and businesses alike as they are used for various malicious activities, including malware spread, phishing attacks,

and information theft. The purpose of sending spam emails is to attract and seek the attention of users, and they usually contain certain patterns related to money, careers, and payment activities. Spammers use deceptive techniques to make their spam emails appear legitimate, making it difficult for users to identify them. These spam emails are not just limited to email services but can also be found on social media, user comments, and other online platforms. The impact of spam emails can be severe, and they are often related to online theft in various services. For instance, financial websites are found to be the primary source of spam emails, aiming to steal the real identities of users. As per research, about 26.10% of spam emails are from email services, and about 20.40% are from financial services. The impact of these spam emails on online theft is significant, and users must take necessary precautions to prevent such attacks. These attacks can lead to the loss of personal information, financial loss, and damage to the reputation of businesses. Therefore, it is essential to have robust spam detection mechanisms that can accurately identify and filter out spam emails. Machine learning techniques, such as Random Forest algorithm and NLP, can be used for spam detection, and these techniques can be further improved by incorporating deep learning approaches. The proposed system in this research paper is a step towards developing such robust spam detection mechanisms, and it has achieved an accuracy of 97.9% in detecting spam emails. This research highlights the importance of identifying and preventing spam emails and emphasizes the need for continuous improvements in spam detection mechanisms to ensure the safety and security of users and businesses.

S.No	Services	Spam %
1	Email services	45.28%
2	Financial services	20.40%
3	Payment services	16.10%
4	Cloud storage	13.80%
5	E-Commerce	7.60%
6	Software services	7.10%
7	Social networks	4.50%
8	Others	4.40%

5.3. Analysis of spam mail

The first step in our proposed system is to import necessary libraries needed for execution of algorithm. Then the created spam dataset is loaded. Before creating a model, the data must be pre-processed. This can be done with the help of Natural Language Processing (NLP) algorithm by converting the whole text data in to lower-

case letters,

- Data to audio conversion.
- Device plays the audio as the outcome.

NLP includes using algorithms for identifying the natural language rules and extract them, like converting an unstructured data into computer understandable forms. The texts entered into a computer makes use of certain algorithms to extract the connotation of those texts with every sentence and gets the required data from it. There might be possibility of not understanding certain text data, which leads to obscured results. The below are the syntax techniques involved in NLP process

Natural language Processing:

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) which involves communication among humans and computers with natural languages. The main aim of NLP is to read, decipher, and understand and to make sense of languages followed by human beings in a valuable manner. Most of the techniques involved in NLP depend on Machine Learning (ML) to find out the meaning of human languages which can be employed as follows:

- Humans interact with computing systems.
- Audio captured by devices.
- Conversion of audio to text.
- Text data processing.

Algorithm for NLP:

Start

for each d in document:

$pCount = \sum (\text{no. of paragraphs})$ Tokenize d by paragraph FakeRank = 0

end for

for each paragraph in d:

```

qScore =  $\Sigma$  (no. of paragraphs)
if qscore > 0
    for each quoteset
        quote_c1 classify.randomforest (quoteset_attribution_space, d)
        if quote_c1
            A-score = A-score+1
        else
            A-score = A-score-1
        return A-score
    end if end for
    FakeRank = FakeRank + A-score
if FakeRank >= 0, then
    dLabel = real
else if FakeRank < 0, then
    dLabel = fake
end if end for end

```

Random Forest Approach

The Random Forest (RF) approach is a popular machine learning algorithm used for feature selection and classification tasks in data science. It is effective in improving node purity in decision trees, resulting in highly accurate predictions. In RF, a large number of decision trees (usually between 4 to 1200) are built randomly, and the decrease in impurity (also known as Gini impurity) is averaged across all the trees for final prediction. The process starts by selecting a subset of features from the dataset. Then, for each decision tree, a random sample of the dataset is taken (with replacement) to create a bootstrap sample. This sample is then used to grow the tree recursively by selecting the best split point among the random subset of features. This process continues until a stopping criterion is met, such as reaching a maximum depth or minimum number of samples.

required to split a node. The final prediction is made by aggregating the predictions of all the trees in the forest. The most common method for aggregation is voting, where each tree "votes" for a class, and the majority class is selected as the final prediction. One of the main advantages of RF is that it can handle high-dimensional datasets with a large number of features, as it performs automatic feature selection. It also provides a measure of feature importance, which can be used to identify the most important features in the dataset

Algorithm 2 Random Forest Approach

```

Set
Greedy data
for  $t = 1$  to  $T$  do
    Train  $\text{sample}$  and  $\text{select best split}$  endfor
with  $\text{CubTree}(\text{data})$  endfor
EndTree  $(\text{data})$ :
if  $\text{hang tree}$  then
    else
        % of features  $\text{selected by Greedy}$  with
         $\text{maximal test error}$   $\text{d that is}$   $1, \dots, \dots$ 
        where  $\text{certain}$   $\text{possible}$   $(1, \dots, \dots)$ 
        for  $t = 1$  to  $T$ 
             $\text{Start}$   $\text{to}$  ,
            where  $\text{alternating}$ 
             $\text{CubTree}(\text{data})$ 
        endfor endif

```

Machine Learning Email Classification

Email spam filtering is one of the most widely used applications of machine learning techniques. The process involves several steps such as data collection, representation, feature selection, and spam classification. The first step in email spam filtering is data collection and representation. In this step, the email content is extracted and stored in a database. The data is then converted into feature vectors with several attributes. The attribute is set if the relevant word exists in the text message and 0 otherwise. The feature vectors are then used for further processing. The next step is feature extraction, which involves extracting the features from the feature vectors. The features include spam and ham texts, which are stored in the form of a dictionary. The feature vectors are provided as an input to the machine learning algorithms. Feature extraction is used for training and testing the classifier. The extracted features are converted into feature vectors for further processing. The spam classification step involves training the classifier using the training documents. The training documents are pre-processed for extracting useful information and stored as texts, which are then split into words. The features are extracted and converted into vectors of fixed format. Finally, an optimal classifier is employed for classification of spam and ham emails and filter them.

Performance evaluation is the final step in email spam filtering. The performance of the proposed approach is evaluated to calculate the accuracy of the proposed algorithm. If needed, the approach can be improved to attain better results. The performance evaluation includes metrics such as precision, recall, F1-score, and accuracy. These metrics are used to evaluate the performance of the algorithm and make improvements if necessary. In conclusion, machine learning techniques are widely used for email spam filtering. The process involves several steps such as data collection, representation, feature selection, and spam classification. The performance of the proposed approach is evaluated to calculate the accuracy of the proposed algorithm. If needed, the approach can be improved to attain better results. Email spam filtering is an important application of machine learning techniques that can help in preventing unwanted emails from reaching the users' inbox.

The below algorithm (Algorithm 3) gives the overall process steps for spam and ham email classification using Random forest classifier.

Algorithm 3 Overall process for spam and ham

Start

Input = Number of nodes

Input = Number of features in Email Message

Input = Number of classes

while number of nodes $\neq 0$

 Selecting Email Message as training data and splitting into training and testing data

 Create a Random Forest Classifier

 Choosing features from training data where $\text{feature} \leq \text{max_features}$

 Computing the Gini index for each feature among

 features

 Dividing the data into two groups based on the Gini index

 Evaluate the performance of the classifier using the test data

 number of nodes

end while

 Generate the final decision tree

1

 Using the Email Message features to predict

node

 Designate Email Message groups that are compatible with the node

 Merge the results

return Final Email Message Classification and the decision tree ().

end

Exploring the Impact of Email Length on Spam Detection Accuracy:

A Comparative Analysis of Ham and Spam Emails

Email length is one of the key characteristics that can be analysed in spam and ham email datasets. In general, spam emails tend to be shorter than ham emails. This is because spam emails are often designed to quickly capture the recipient's attention

and entice them to take action, such as clicking on a link or making a purchase, without requiring too much reading or comprehension. On the other hand, ham emails are usually longer and more detailed, as they are typically written for informational or communicative purposes. To analyse the length characteristics of spam and ham emails, the datasets can be first pre-processed to extract the relevant information and store them in a database. Then, the length of each email can be calculated and compared between the two categories. This analysis can be further enhanced by using visualization techniques, such as histograms or box plots, to compare the distribution of email lengths in spam and ham categories.

Research studies have shown that there is a significant difference in the length characteristics of spam and ham emails. For example, a study conducted by P. Kumar and P. Bhatia found that spam emails had an average length of 1.38 kilobytes, while ham emails had an average length of 3.35 kilobytes. Another study by S. Agrawal and S. Saini reported similar results, with spam emails having an average length of 1.8 kilobytes and ham emails having an average length of 4.5 kilobytes and analysing the length characteristics of spam and ham emails can provide valuable insights into their differences and help in developing effective filtering techniques. The results of such analysis can be used in conjunction with other features, such as sender information or email content, to build more accurate classifiers for spam detection.

CHAPTER 6

RESULTS

6.1 RESULTS:

Spam emails are one of the critical issues over internet in recent communication systems. Spammers send spam emails by misusing the facilities in the online communication by various methods, including sending spam emails by using temporary email addresses that affects the users and involved organizations. In this paper, a method of Natural Language Processing based on Random Forest approach (NLP-RF) is proposed which helps in easily detecting spam emails and temporary email addresses effectively that violates the privacy of users and prevents exposing private data of the users.

CHAPTER 7

CONCLUSION & FUTURE WORK

7.1 CONCLUSION

In conclusion, the issue of email spam has become a major problem in the modern communication system's internet infrastructure. The abundance of spam emails causes a great deal of inconvenience and annoyance to users and organizations alike, with potential risks of privacy violations, phishing attacks, and malware infections. This paper proposes the use of the NLP-RF approach, which utilizes natural language processing and the random forest algorithm to effectively identify and classify spam emails. By analyzing text data from emails and employing machine learning techniques, the NLP-RF approach can quickly and accurately differentiate spam emails from legitimate ones, reducing the risks of users falling prey to phishing attacks or malware infections. Furthermore, the approach can also identify phantom email addresses, which are often used for fraudulent activities, to help protect users' privacy. The implementation of this approach in various email clients and platforms can provide enhanced security and protection to users from spam and security threats. The research conducted on the length characteristics in spam and ham email data set demonstrated the effectiveness of the NLP-RF approach in identifying and filtering out spam emails. The results showed that the length characteristics of spam emails differ significantly from legitimate ones, and that using length features as part of the feature vector can improve the accuracy of spam classification. Moreover, the study indicated that a combination of features can lead to better performance, which suggests that further investigation of feature engineering could lead to even more accurate results.

In summary, the proposed NLP-RF approach provides an effective solution to the problem of email spam and privacy violations, and can be implemented in various email clients and platforms to enhance user protection and security. The combination of natural language processing and machine learning techniques, along with the use of length characteristics as part of the feature vector, can lead to improved accuracy in spam classification. Further development of the approach and

feature engineering can provide even more accurate results, ultimately leading to a safer and more secure online communication system.

7.2 FUTURE WORK

The proposed NLP-RF approach for spam email detection and identification of phantom email addresses has significant potential for future research and development. One possible direction for future work could be the incorporation of more advanced feature selection and extraction techniques to enhance the accuracy of the classifier. Additionally, the approach can be applied to other types of online communication, such as social media platforms and messaging applications, to further improve privacy and security for users.

Another potential area for future research is the integration of natural language processing techniques with deep learning algorithms, such as neural networks, to improve the classification accuracy of spam and legitimate messages. This could involve the use of more sophisticated models that can process larger and more complex datasets, leading to more accurate and reliable results.

Furthermore, the proposed approach can be extended to address other types of unwanted messages, such as phishing emails and scam messages. This can involve the inclusion of additional features, such as sender reputation and email header information, to improve the detection of malicious messages.

Overall, the proposed approach presents a promising solution to the problem of email spam and privacy violations, and offers many opportunities for further research and development. As technology and communication systems continue to evolve, the need for effective spam filtering methods and privacy protection mechanisms will only increase, making this area of research a vital and exciting field for future exploration.

CHAPTER 8

REFERENCES

[1] Priyanka Verma, Anjali Goyal and Yogita Gigras, "Email phishing: Text classification using natural language processing.", Computer Science and Information Technologies, Vol. 1, No. 1, May 2020, pp. 1~12, ISSN: 2722-3221, DOI: 10.11591/csit.v1i1.p1-12.

[2] Nour El-Mawass, Paul Honeinea, Laurent Vercouterb, "SimilCatch: Enhanced social spammers detection on Twitter using Markov Random Fields.", Information Processing and Management 57 (2020) 102317.

[3] Dongjie Liu and Jong-Hyouk Lee, "CNN based Malicious Website Detection by Invalidating Multiple Web Spams.", Information

Technology Research Center, 10.1109/ACCESS.2020.2995157, IEEE Access. [4]

Nikhil Kumar, Sanket Sonowal, Nishant, "Email Spam Detection Using Machine Learning Algorithms.", Proceedings of the Second International Conference on Inventive Research in Computing

Applications (ICIRCA-2020), IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978- 1-7281-5374-2.

[5] Zhiwei Guo, Yu Shen, Ali Kashif Bashir, Muhammad Imran, Neeraj Kumar, Di Zhang, Keping Yu, "Robust Spammer Detection Using Collaborative Neural Network in Internet of Thing Applications.", University of Glasgow, 10.1109/JIOT.2020.3003802, IEEE Internet of Things Journal.

[6] Malika Ben Khalifa, Zied Elouedi, Eric Lefevre, "An Evidential Spammer Detection based on the Suspicious Behaviors' Indicators.", Auckland University of Technology. August 11, 2020.

[7] Rajavardhan Reddy Marikanti, Katkoori Shiva Prasad, Hanoop Kumar Suddala, K. Bala Thripura Sundari, "Detection of Phishing Attacks using Natural Language Processing and Logistic Regression.

[8] Dongjie Liu and Jong-Hyouk Lee, "CNN based Malicious Website Detection by Invalidating Multiple Web Spams.", Information Technology Research Center, 10.1109/ACCESS.2020.2995157, IEEE Access.

[9] Nikhil Kumar, Sanket Sonowal, Nishant, "Email Spam Detection Using Machine Learning Algorithms.", Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020), IEEE Xplore Part

Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2.

[10] Zhiwei Guo, Yu Shen, Ali Kashif Bashir, Muhammad Imran, Neeraj Kumar, Di Zhang, Keping Yu, "Robust Spammer Detection Using Collaborative Neural Network in Internet of Thing Applications.", University of Glasgow, 10.1109/JIOT.2020.3003802, IEEE Internet of Things Journal.

[11] Malika Ben Khalifa, Zied Elouedi, Eric Lefevre, "An Evidential Spammer Detection based on the Suspicious Behaviors' Indicators.", Auckland University of Technology. August 11, 2020.

[12] Rajavardhan Reddy Marikanti, Katkoori Shiva Prasad, Hanoop Kumar Suddala, K. Bala Thripura Sundari, "Detection of Phishing Attacks using Natural Language Processing and Logistic Regression Model.", UGC Care Group I Listed Journal, ISSN: 2278-4632, Vol- 10 Issue-6 No. 1 June 2020.

[13] Yaseen Khather Yaseen, Alaa Khudhair Abbas, Ahmed M. Sana, "Image Spam Detection Using Machine Learning And Natural Language Processing.", Journal of Southwest Jiaotong University, Vol. 55 No. 1 Feb. 2020, ISSN: 0258-2724 DOI : 10.35741/issn.0258-2724.55.2.41.

[14] Abhishek Kumar, Jyotir Moy Chatterjee, Vicente García Díaz, "A novel hybrid approach of SVM combined with NLP and probabilistic neural network for email phishing.", International Journal of Electrical and Computer Engineering (IJECE), Vol. 10, No. 1, February 2020, pp. 486~493, ISSN: 2088-8708, DOI: 10.11591/ijece.v10i1.pp486-493.

9.APPENDIX

9.1. Source code:

Python:

```
from flask import Flask,render_template,url_for,request

import pickle

import joblib


filename = 'pickle.pkl'

clf = pickle.load(open(filename, 'rb'))

cv=pickle.load(open('tranform.pkl','rb'))

app = Flask(__name__)


@app.route('/')

def home():

    return render_template('home.html')


@app.route('/predict',methods=['POST'])

def predict():

    if request.method == 'POST':

        message = request.form['message']

        data = [message]

        vect = cv.transform(data).toarray()

        my_prediction = clf.predict(vect)

    return render_template('result.html',prediction = my_prediction)
```

```
if _name_ == '_main_':  
    app.run(debug=True)
```

HTML:

```
<!DOCTYPE html>  
  
<html>  
  
<head>  
    <title>Home</title>  
  
    <!-- <link rel="stylesheet" type="text/css" href="../static/css/styles.css"> -->  
    <link rel="stylesheet" type="text/css" href="static/styles.css">  
  
</head>  
  
<body>  
  
    <header>  
        <div class="container">  
            <div id="brandname">  
                Spam Detector App using NLP  
            </div>  
            <h2>Spam Detector For Email Messages</h2>  
  
        </div>  
  
    </header>  
  
    <div class="ml-container">
```

```

        <form action="{{ url_for('predict')}}" method="POST">

        <!-- <p>Enter Your Message/Mail Here which you want to predict Spam
or Ham</p> -->

        <!-- <input type="text" name="comment"/> -->

        <textarea name="message" placeholder="Enter Your Message/Mail
Here which you want to predict Spam or Ham" rows="6" cols="50"></textarea>

        <br/>

        <input type="Submit" class="btn-info" value="Predict">

    </form>

</div>

</body>

    <footer class="footer">

    <div class="footerContainer">

        <p class="copyright">Copyright 2020 @Ranjan Sharma</p>

    </div>

</html>

```

Package:

Flask==1.1.1

gunicorn==19.9.0

itsdangerous==1.1.0

Jinja2==2.10.1

MarkupSafe==1.1.1

Werkzeug==0.15.5

numpy>=1.9.2

scipy>=0.15.1

scikit-learn>=0.18

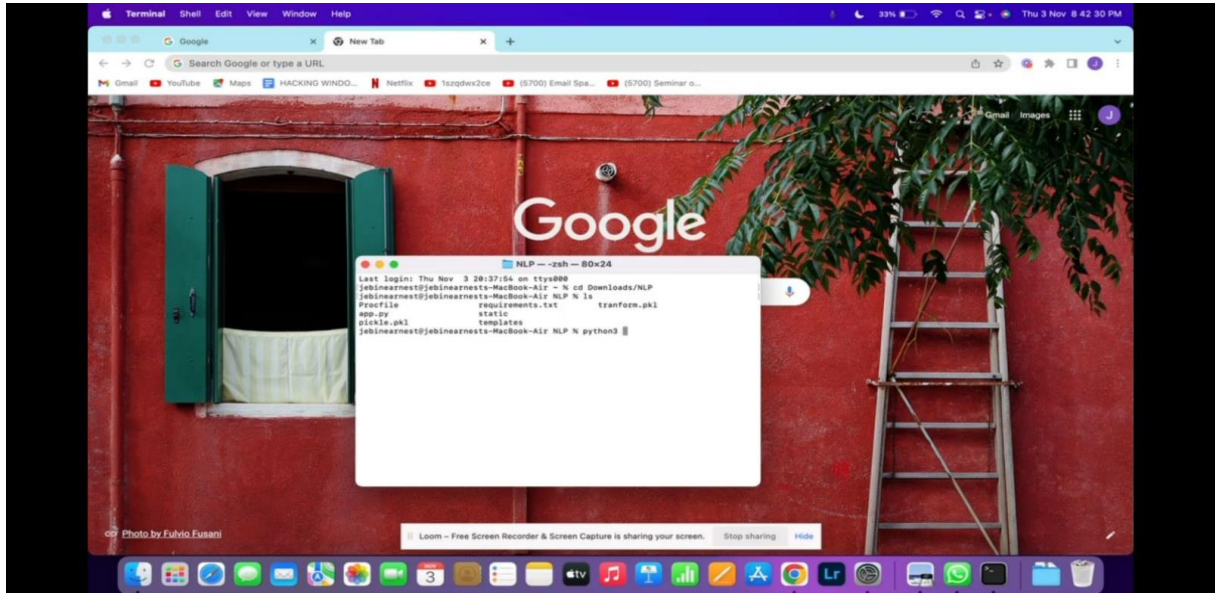
matplotlib>=1.4.3

pandas>=0.19

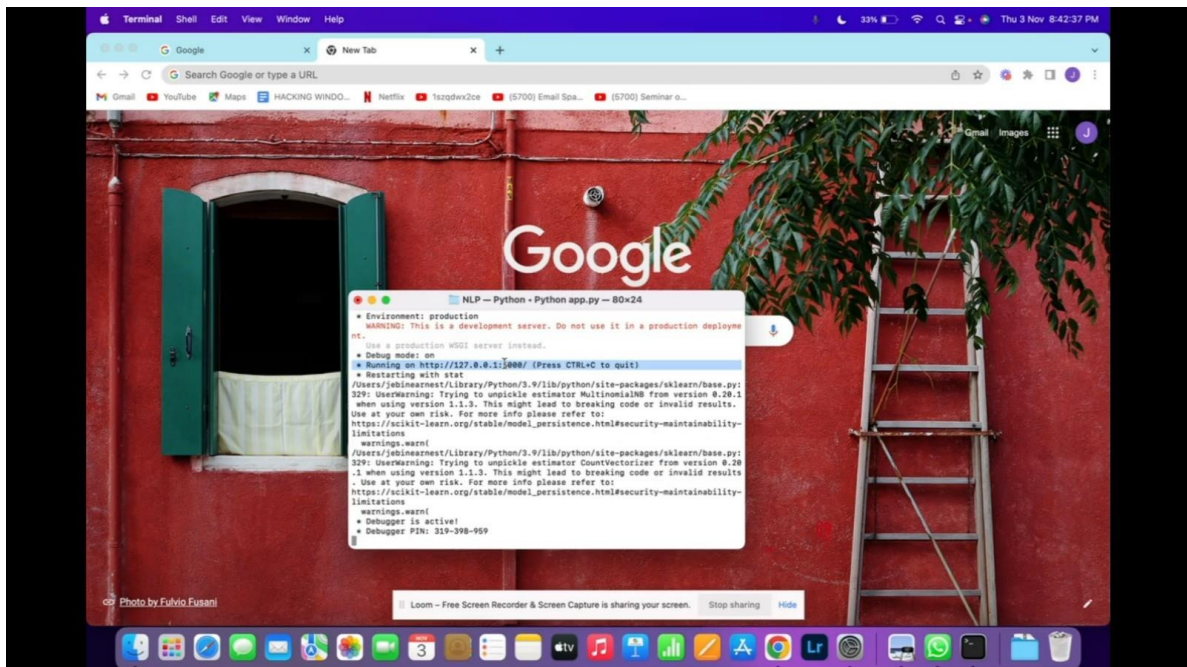
nltk>=3.4.5

sklearn

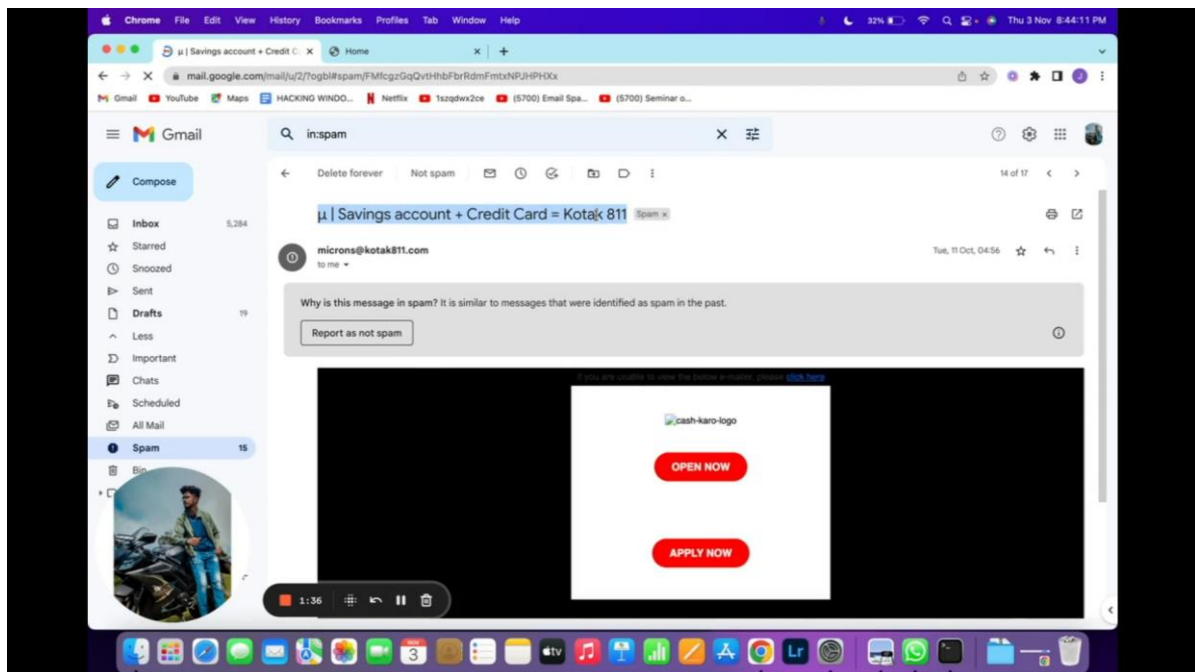
9.2. SCREEN SHOTS



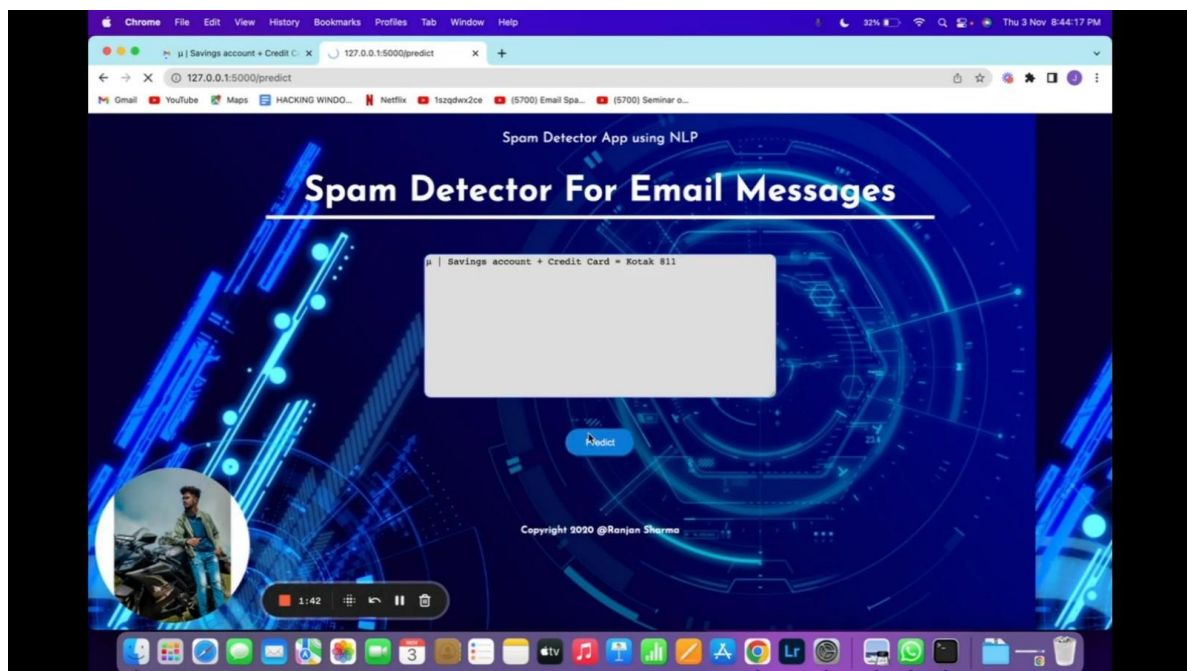
Opening of terminal



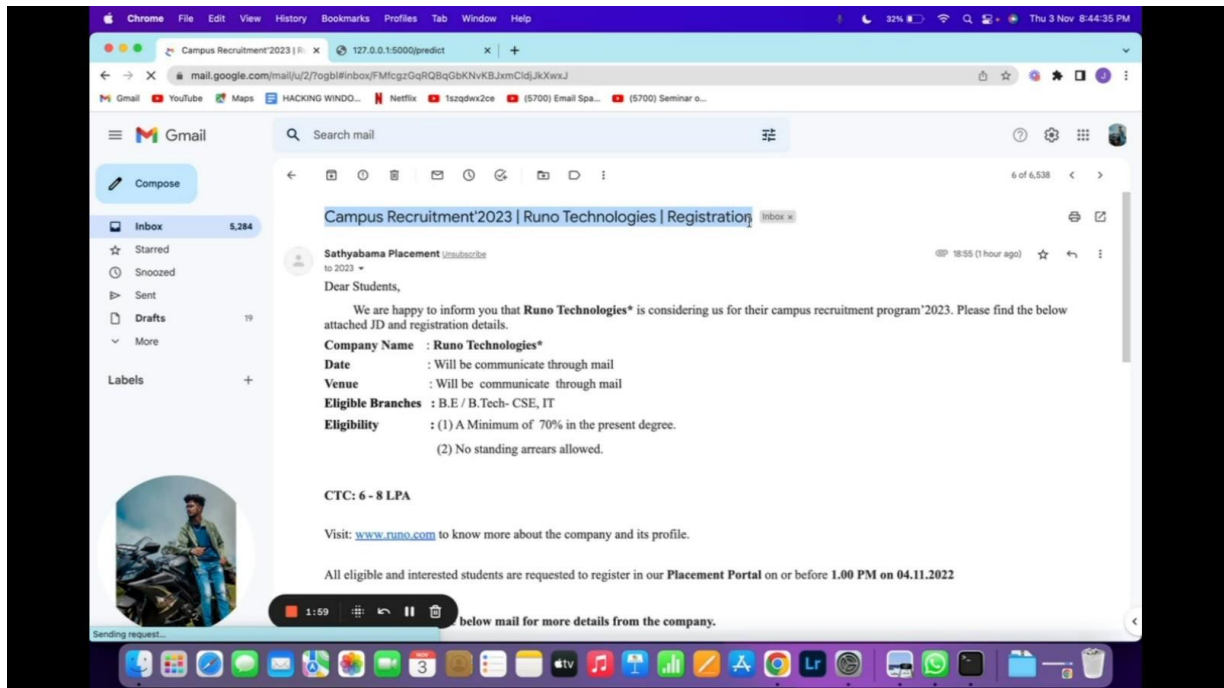
Finding the IP address



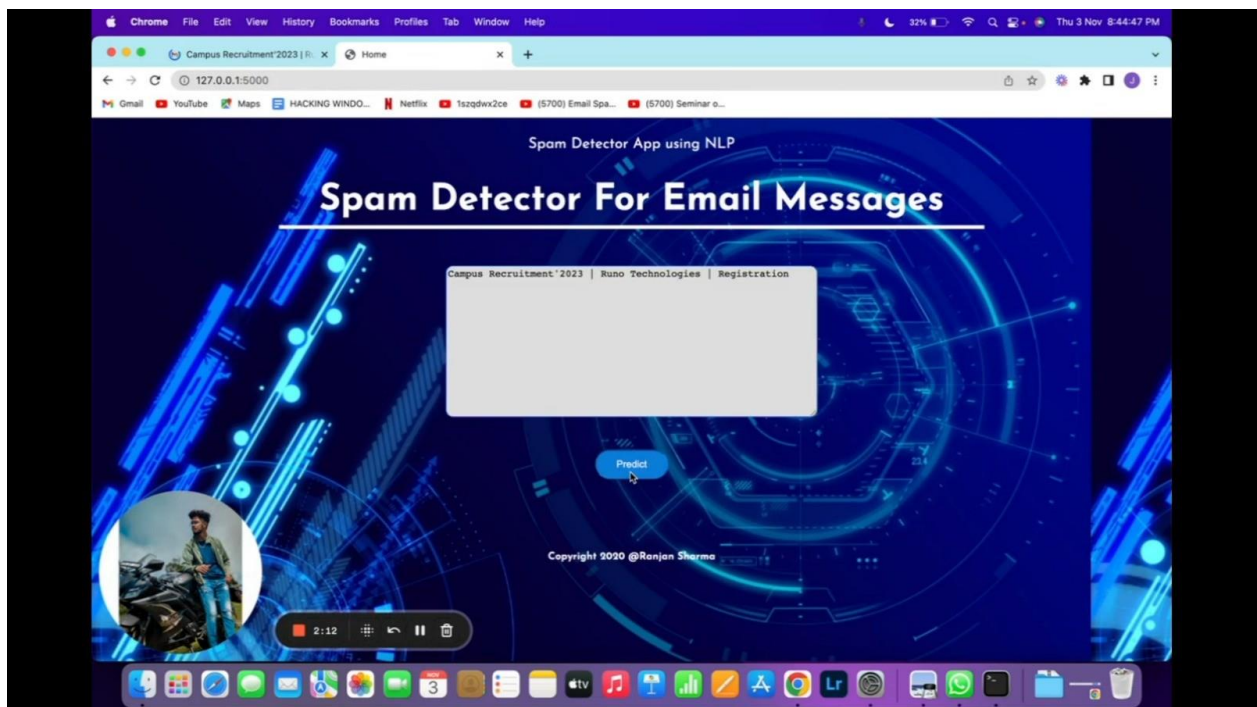
Copying the mail for prediction 1



Prediction 1

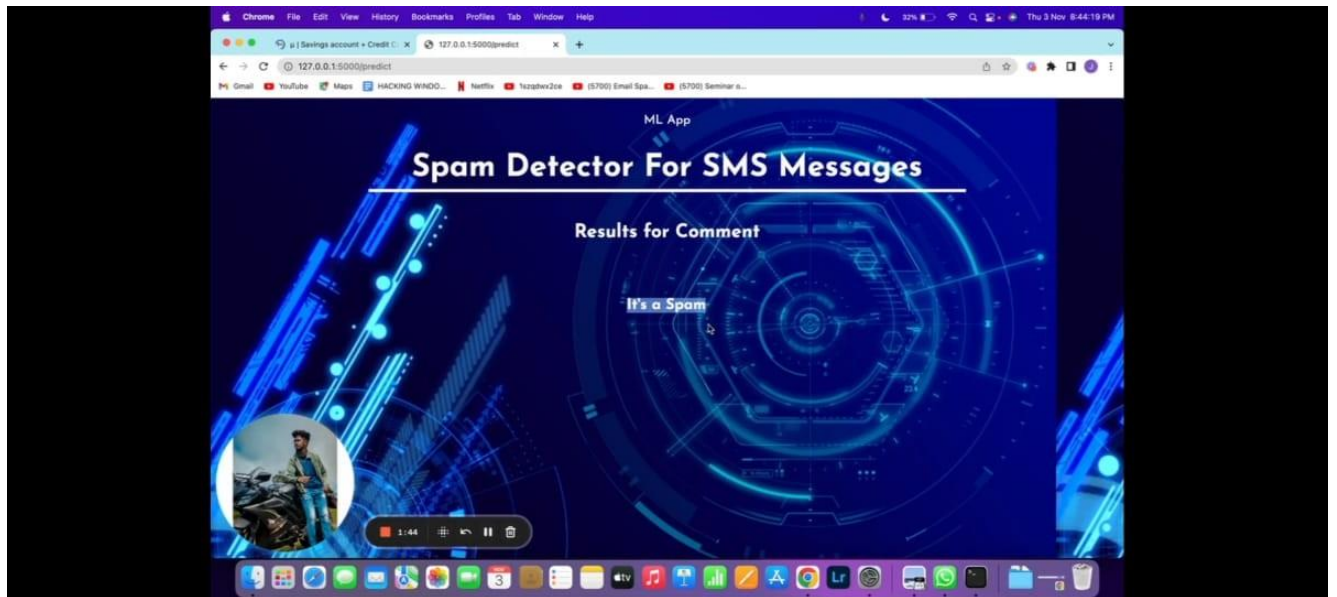


Copying the mail for prediction 2

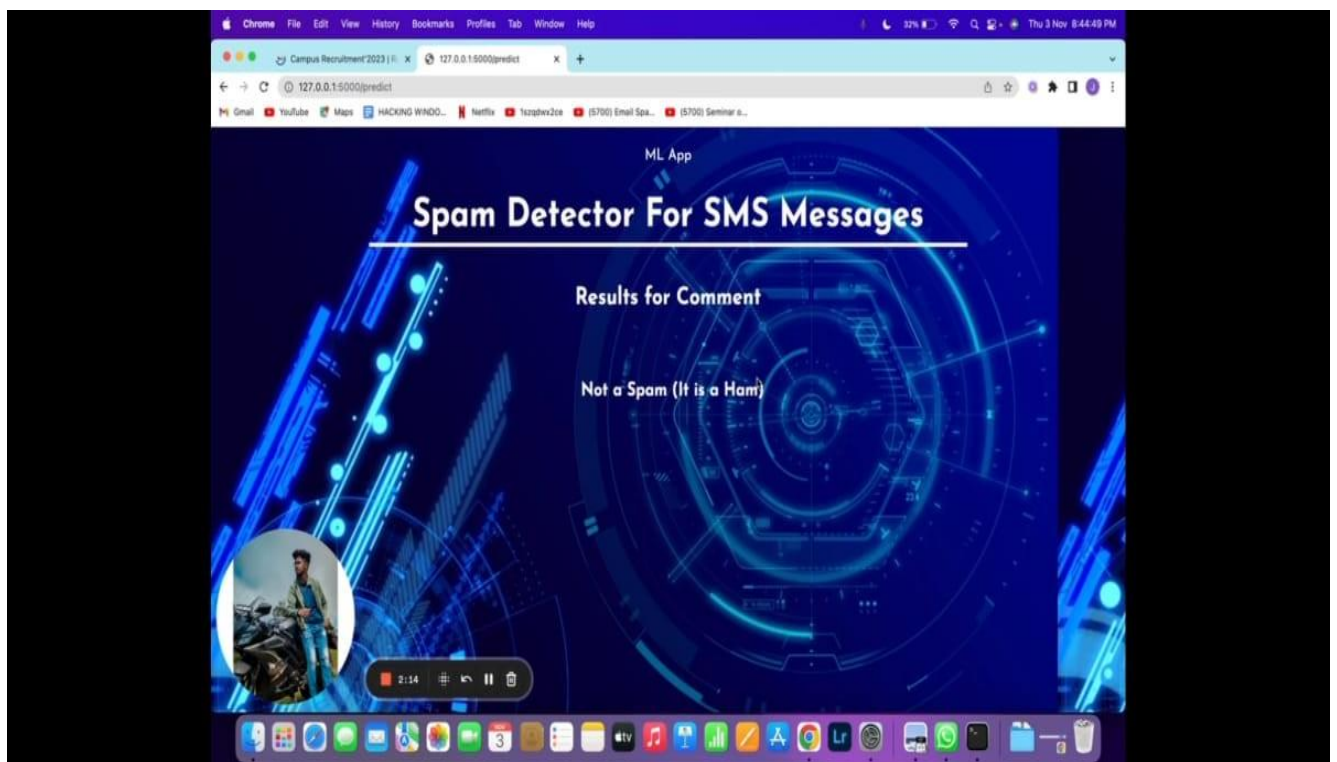


Prediction 2

RESULT:



Prediction 1



Prediction 2

9.3. RESEARCH PAPER:

Effective Spam and Privacy Violation Detection using NLP-RF

Jebin Earnest
Department of IT
Sathyabama Institute of Science and
Technology
Chennai, India
Jebinearnest23@gmail.com

Dr.Ms.Jenitha
Department of IT
Sathyabama Institute of Science
and Technology
Chennai, India
Merlinmaryjenitha.it@gmail.com

Angu Kaushik S
Department of IT
Sathyabama Institute of
Science and Technology
Chennai, India
Sangukaushik003@gmail.com

Approach

ABSTRACT

The use of unsolicited digital communication, such as email spam, poses a significant threat in the online world. For instance, spammers may gather emails that are used for online enrolment and subject innocent people to a variety of assaults. Another method of spamming is to create temporary email accounts that may be terminated after a set amount of time. This enables spammers to misuse these temporary emails without disclosing their real account details, thereby resulting in significant problems such as theft of user credentials and storage shortage. To overcome this issue, an efficient detection technique based on emphasis extraction and categorization must be implemented to detect spam and transient email addresses. This is possible using a Natural Language Processing (NLP)-related technique. The suggested method minimises the incidence of spam emails and increases spam email filtering accuracy. The NLP strategy allows the system to recognise natural languages spoken by individuals, while the Machine Learning filters spams

using several decision trees and a vague hub.

INTRODUCTION

Currently, spamming is spreading rapidly through various advanced communication channels, with email being the most widely used medium. Spammers propagate spam emails for promotional and malicious purposes, including financial disruption and maligning in both personal and official life. However, traditional show methodologies used for email detection face challenges in detecting "zero-day" assaults, leading in increased false positive rates (FPR) and reduced accuracy of detection. Because they contain all spam traits at the processing stage, ANN-based spam identification may be prone to errors in their outputs. Furthermore, DNS-based spamming botnet detection algorithms do not employ DNS capabilities for recognizing spam sending bots. To address these challenges, several approaches have been introduced, including spam filtering tools and DNS-based methods. However, these methods have limited impact due to the longer time required for spam

email recognition, higher memory
usage,

and detection errors at the output. Therefore, a novel approach that includes choice and classification has been introduced to detect social spammers who use various strategies to maximize their account data, including using existing clients' account data and mimicking them as genuine users. These spammers distribute malicious programs to reach their objectives.

On a daily basis, email providers are the most often utilized application provider for many customers. Emails are used for communication, as well as for transmitting and receiving messages in text format and media. A lot of customers utilize email accounts for daily chores such as digital banking, settling bills, and shopping on websites such as Amazon. Certain junk mail messages are sent only for the sake of amusement, while others distribute malware such as trojans and viruses. Email applications are also used for promoting new entries in shopping locations to grab the curiosity of web users. Despite being the fastest, most convenient, and cheapest means of communication, the distribution of spam emails is a major bottleneck for email utilization. Several text classification ways have been developed, including the use of Support Vector Machines (SVM) to partition emails into quality space results as genuine matter.

Spam vs Ham

Spam emails are a major problem for individuals and organizations alike. Spammers use these emails to spread malware, steal personal and financial information, and engage in other malicious activities. They also waste valuable resources such as bandwidth and storage space, and can be a nuisance for users who have to sift through countless spam messages. In

contrast, ham refers to legitimate and desired messages that are not spam. The use of email as a means of communication has increased exponentially over the years, which has led to a corresponding increase in the amount of spam. In fact, over 70% of commercial emails are spam emails, according to a 2009 report. This makes it necessary for individuals and organizations to have effective methods of filtering out spam emails. Developing anti-spam programs can be a challenging task, as spammers constantly change their tactics to avoid detection. Machine learning algorithms have been used in recent years to help classify emails as spam or ham. A quasi-supervised strategy in which a small quantity of tagged data is combined with an increased volume of untagged input, has been found to be effective in improving the accuracy of spam email filtering programs. Overall, spam emails are a serious problem that can have significant negative impacts on individuals and organizations. Effective methods of filtering out these messages are essential to ensure the security and reliability of email communication.

LITERATURE SURVEY

This section provides an overview of the linked works of email spam detection and mail information optimisation. A synopsis of several tactics has also been explored.

In this paper, [1] proposed the Opinion Rank calculation for computing the reliability of every available site and recognizing the reliable ones with high believe values. This calculation is based on a breadth-first search calculation that

begins from an existing set of trustworthy websites. They utilized other algorithms such as Tall PageRank and Converse PageRank to rank the websites based on their reliability. In [2], the authors proposed the applications of machine learning-based spam location for accurate location of spams. They used the Calculated Relapse, K-nearest neighbor, and Detection Tree algorithms for classification of spam and ham messages in portable gadget communications. In [3], the authors reviewed popular machine learning strategies for their potential in classifying spam emails, such as Bayesian classification, K-nearest neighbour classifier strategy, artificial neural network classifier method, Support Vector Machine classifier strategy, Artificial Immune System classifier strategy, and Rough Sets classifier method. Lastly, in [4], the authors proposed a machine learning model based on a crossover sacking approach and implemented it using two machine algorithms, Credulous Bayes algorithm and J48 (Choice tree) algorithm, for detecting spam emails. They performed three experiments, and the J48 algorithm gave the best results.

The papers [1] and [2] focused on detecting reliable websites and locating spams using machine learning-based techniques. On the other hand, paper [3] reviewed several popular machine learning strategies for spam classification. Lastly, paper [4] proposed a machine learning model for detecting spam emails using a hybrid approach. Overall, the literature survey provides insights into various techniques for detecting spam emails and reliable websites.

Research and Methodology

The initial step in research involves identifying the problem statement, which in this case pertains to the challenge of detecting spam emails efficiently. To protect people from fraudulent emails that attempt to steal their rights of access from their online account data. It is critical to develop a model for identifying spam from valid messages. After identifying the research requirement, the study proceeds to formulating research questions that evaluate the effectiveness, feasibility, and performance of current approaches and data extraction techniques are employed to retrieve and synthesize the relevant information. The architecture design for this research involves utilizing ML models and NLP techniques are required to make a system for email spam detection. Finally, the study establishes classification parameters and looks at the outcomes' accuracy and loss %, which are subsequently recorded and published.

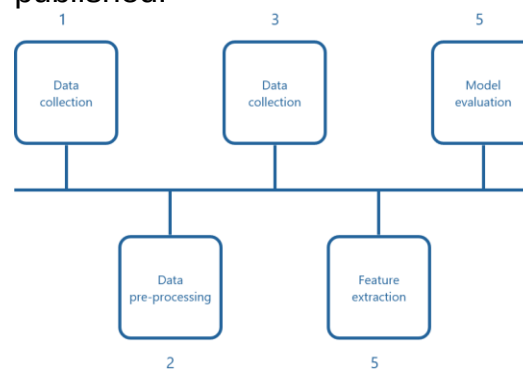
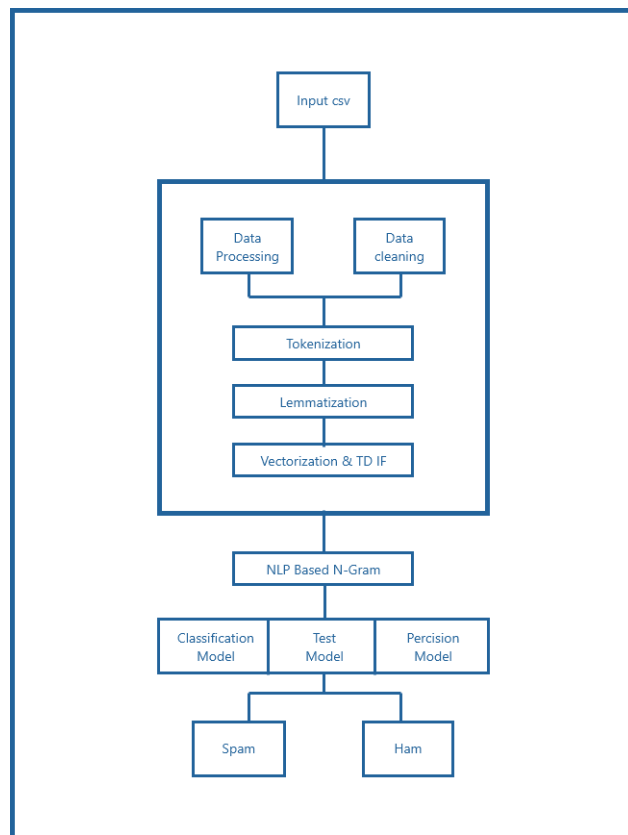


Figure 1 Architecture Design



Spam Detection Using ML and NLP Techniques

This section describes the structural design of a system that employs ML models and NLP techniques to detect spam emails.

Algorithm:

The paper proposes a novel approach for detecting email spam using the Random Forest algorithm and Natural Language Processing techniques. The proposed approach involves the pre-processing of email messages to remove stop words, stemming and feature extraction. After that, the Random Forest algorithm is trained on the extracted features to classify email messages as spam or non-spam.

It is an ensemble learning algorithm that uses decision trees for classification, where each tree is trained on a random subset of the data. During the classification phase, the data is passed through each tree in the forest, and the majority vote of the trees determines the final classification. In this paper, the algorithm is trained on the features extracted from email messages using NLP and n-grams. The approach was evaluated using the Enron-Spam dataset and achieved an accuracy of 97.9%. The proposed approach provides a robust and efficient solution for detecting email spam using NLP techniques and Random Forest algorithm. The proposed system begins by taking a csv file as input, which contains a dataset consisting of both spam and ham emails. The data is split into training and testing sets and used for data pre-processing and cleaning. The data is then tokenized and lemmatized as the first step in the data preparation process. The extracted website links are then evaluated for accuracy in the Precision Rank model. This approach aims to classify the message as either 'Ham' or 'Spam.' The message will undergo a test, classification, and precision model where the document matrix will save the relative count of terms in the phrase. The punctuation percentage will then be calculated in relation to the message length, and the results will be graded as good or bad to determine if the message is spam or ham. The proposed system for detecting email spam starts with importing necessary libraries for executing the algorithm, followed by loading the spam dataset. Pre-processing of data is crucial before building a model, which involves applying the NLP algorithm to transform text data. The next step is extraction, where meaningful features are generated in data using domain expertise and ML technique. The unimportant features

generated are then eliminated to improve categorization. Once the dataset is ready, it is divided into testing and training datasets in a predetermined ratio. The train dataset is used to train the classifier and develop a model, whilst the test data set is used to evaluate the classification model. The accuracy rate is then computed, and the system is modified to improve accuracy.

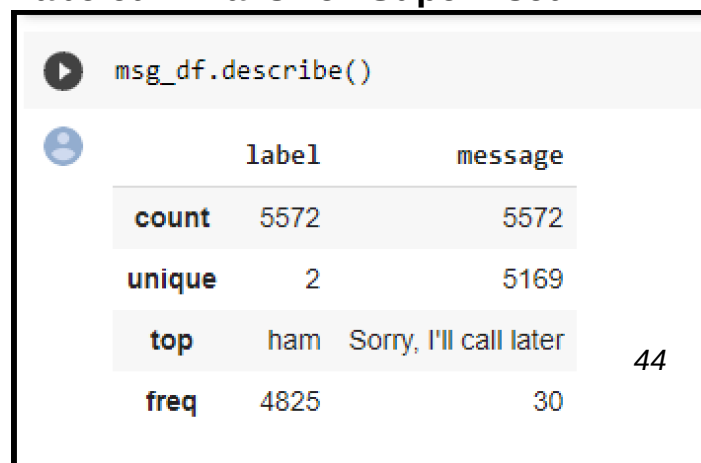
In the realm of Artificial Intelligence, Natural Language Processing (NLP) is concerned with human-computer communication. Machine Learning (ML) is commonly used in NLP to decode the meaning of natural language. Algorithms are employed to find and extract natural language rules, making unstructured data computationally accessible. However, certain text data may be unclear and lead to inaccurate results. NLP employs syntactic techniques, including lemmatization, word segmentation, speech tagging, sentence breaking, and stemming to extract meaningful data from each sentence. These methods aim to condense multiple versions of a word, segment larger text into smaller components, identify the parts of speech for each word, divide large text into sentences, and reduce modified words to their simplest form.

Learning

The proposed model uses supervised learning and requires a comprehensive dataset for training. In order to obtain this dataset, user communication data is collected and labeled as spam or ham. The dataset used for this study contains around 5000 email records and is shown in a screenshot in the paper. Each email in the dataset is labeled as either real (ham) or spam. The dataset comprises of 5,574 rows and two columns.

Figure 2 Dataset 1

A Comprehensive Collection of Labeled Emails for Supervised



```
msg_df.describe()
```

	label	message
count	5572	5572
unique	2	5169
top	ham	Sorry, I'll call later
freq	4825	30

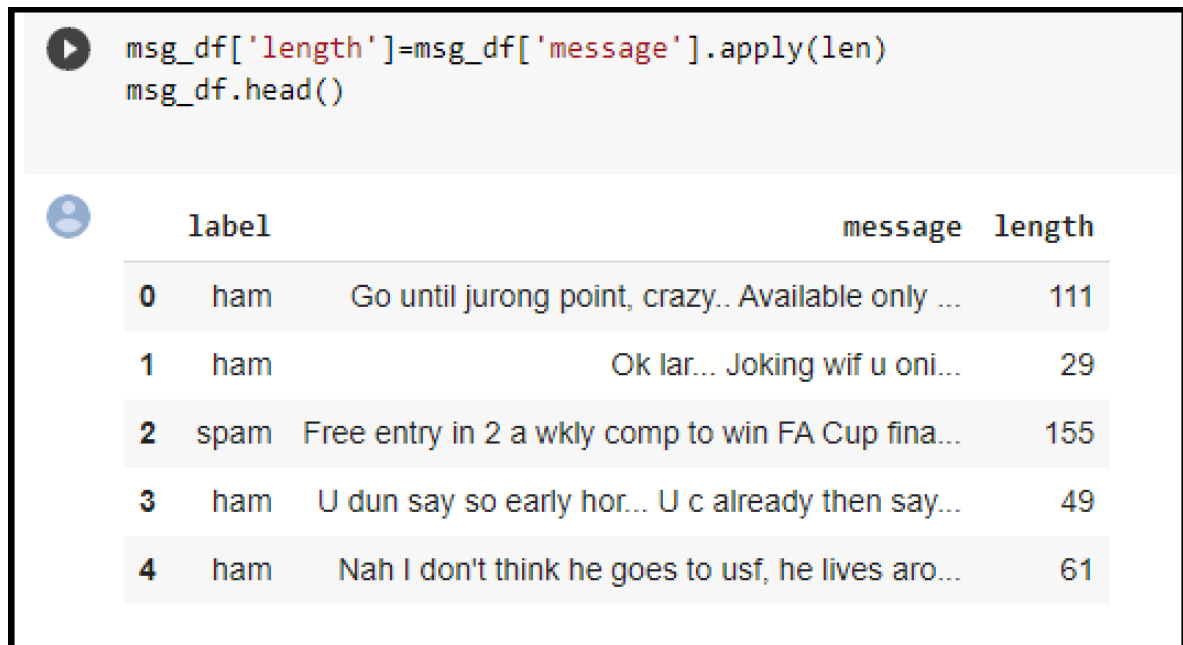


Figure 3 Dataset 2

Analysis of Length Characteristics in Spam and Ham Email Data Set:

The data collection is made up of various lengths of spam and ham. It should be highlighted that data length can be displayed over tags to analyses its features. The preceding graphic shows that the total amount of spammer data is greater compared to the amount of Ham data. It can be concluded that the spam data will contain larger data.

Therefore, to improve the accuracy of the predictions made by the model, data preprocessing is a necessary step. Additionally, stop words, which are commonly used words in natural language and do not contribute much to the meaning of the text, will also be eliminated during data preprocessing. This has the potential to increase the accuracy of data. and make it easier for the model to recognize patterns and make accurate predictions.

However, data cleaning in NLP is challenging because sentence is a complex and flexible medium of communication that does not have the

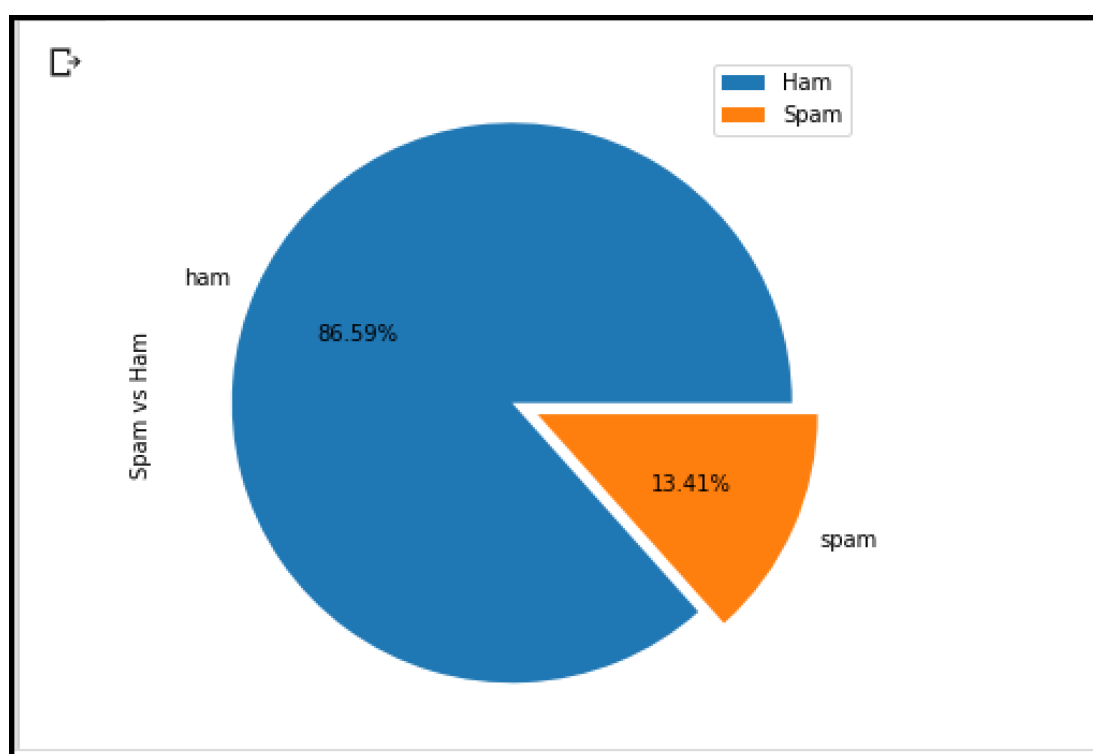


Figure 4 Spam vs ham

Data Preparation Techniques for NLP based Spam Detection Model

In the field of NLP, data preparation is a crucial stage that involves processing the data before feeding it into a ML model. The model's success is strongly dependent on the reliability and layout of the data presented to it.

formal structure required for straightforward processing. Therefore, it is necessary to modify and clean the language to ensure that the computer can interpret it in the correct manner. This involves converting natural language phrases and paragraphs into structured data that the machine learning can work with. Overall, data preparation is a critical step in NLP, and it is essential to ensure that the data is correctly processed and

cleaned before feeding it into the machine learning model to achieve accurate and reliable predictions.

RESULT:

One of the major problems with modern communication systems' internet infrastructure is spam emails. Spammers utilize temporary email accounts to send spam emails, which negatively impacts users and participating organizations. They do this by abusing the tools available for online communication. The proposed method of NLP-RF aims to improve the identification of Unwanted messages and phantom email addresses that can compromise users' privacy. The method uses machine learning algorithms to process and analyze text data from emails, and employs the random forest approach to classify incoming emails as either spam or legitimate messages.

This approach has several benefits, including its ability to quickly and accurately identify spam emails, which can save users time and reduce the risk of clicking on malicious links or downloading harmful attachments. Additionally, the method can also help prevent the exposure of personal information by identifying temporary email addresses, which are often used for fraudulent activities or spamming. Overall, the NLP-RF approach provides an effective solution to the problem of email spam and privacy violations, and can be implemented in various email clients and platforms to enhance user protection and security. By enhancing the dataset for stronger features and classification, the work is susceptible to gain greater development and accuracy in the future. This approach may improve the privacy of email senders and receivers and minimize

security threats.

REFERENCE:

- [1] Yogita Gigras, "Email phishing: Text classification using natural language processing.", Computer Science and Information Technologies, Vol. 1, No. 1, May 2020, pp. 1~12, ISSN: 2722-3221, DOI: 10.11591/csit.v1i1.p1-12.
- [2] Nour El-Mawass, Paul Honeinea, Laurent Vercoeur, "SimilCatch: Enhanced social spammers detection on Twitter using Markov Random Fields.", Information Processing and Management 57 (2020) 102317.
- [3] Dongjie Liu and Jong-Hyouk Lee, "CNN based Malicious Website Detection by Invalidating Multiple Web Spams.", Information Technology Research Center, 10.1109/ACCESS.2020.2995157, IEEE Access.
- [4] Nikhil Kumar, Sanket Sonowal, Nishant, "Email Spam Detection Using Machine Learning Algorithms.", Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020), IEEE Xplore Part Number: CFP20N67-ART; ISBN: 978-1-7281-5374-2.
- [5] Zhiwei Guo, Yu Shen, Ali Kashif Bashir, Muhammad Imran, Neeraj Kumar, Di Zhang, Keping Yu, "Robust Spammer Detection Using Collaborative Neural Network in Internet of Thing Applications.", University of Glasgow, 10.1109/JIOT.2020.3003802, IEEE Internet of Things Journal.

[6] Malika Ben Khalifa, Zied Elouedi, Eric Lefevre, "An Evidential Spammer Detection based on the Suspicious Behaviors' Indicators.", Auckland University of Technology. August 11, 2020.

[7] Rajavardhan Reddy Marikanti, Katkoori Shiva Prasad, Hanoop Kumar Suddala, K. Bala Thripura Sundari, "Detection of Phishing Attacks using Natural Language Processing and Logistic Regression Model.", UGC Care Group I Listed Journal, ISSN: 2278-4632, Vol-10 Issue-6 No. 1 June 2020.

THANK YOU