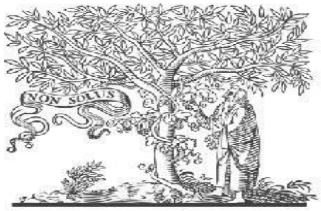
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 31st Mar 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 03

10.48047/IJIEMR/V12/ISSUE 03/111

Title **Email or Message Based Spam Detection Using Machine Learning**

Volume 12, ISSUE 03, Pages: 793-804

Paper Authors
T. Harshini Sai Sree , G. Naveen Saai , M. Abhisheak, Y. Saidarao

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Email or Message Based Spam DetectionUsing Machine Learning

T. Harshini Sai Sree [1], G. Naveen Saai [1], M. Abhisheak[1], Y. Saidarao[1]

UG Students, Dept of CSE, Kallam Haranadhareddy Institute of Technology, Ap, India.

ABSTRACT

This project's goal is to detect suspicious people by blocking emails that include abusive or antisocial content and suspecting their senders. Untrustworthy people are flagged in a type of mailing system called suspicious email detection. determined by figuring out the keywords he/she used. E-mail spoofing is the term for malicious conduct where the origin details are changed to make the email appear to come from a different source. The defense division of any government is where this method is primarily used. The true sender's email is forged at the attacker's end using an SMTP server and an SPF record to make the email appear more legitimate and authentic. Since email spoofing is the most often used weapon for social engineering, it has become a crucial component of every investigation agency and intelligence service.

Keywords: suspicious, spam, phishing.

1. INTRODUCTION

The internet has progressively assimilated into daily life. The number of people using email is growing daily as a result of increased internet usage. Because of the rise in email usage, there are issues brought on by spam, or unsolicited bulk email. Spam emails are becoming common since email has emerged as one of the finest mediums for advertising. Generated.
Spam emails are those that the recipient has requested not to receive. To many email recipients, many copies of the same message are sent. Giving away our email address on an unlawful or dishonest website frequently results in spam. Spam has a wide range of negative impacts. Phishing is viewed as a difficult problem in today's society that is escalating quickly every year. Using social engineering and technical ways to steal customers' private information, such as usernames and passwords, is seen as a crime (Manning & Aron 2015). Lungu and Tabasco contend that in this regard, the current economic crisis is a mirror of the rise in hacking attempts and other invasions of internet users' privacy (Lungu & Tabusca, 2010). According on the applicable channel of proliferation, phishing techniques are divided into different categories. They include malware, phishing emails, and fake websites (Jain & Richariya 2011). Spam messages fall under the category of phishing emails.
Users receive emails inviting them to click on an embedded link that they are told are from a reputable company or bank. The link will take the user to an impersonated website that asks for private information like usernames, passwords, or credit card details (Al-Momani and Gupta 2013).

With advancements in internet technology and the ensuing revolution in online user engagement, security concerns have grown more serious. The user of the internet is threatened by the always changing security challenges, which could result in financial and identity loss. Phishing is a type of social engineering threat that takes advantage of the ignorance of uneducated internet users to trick them into giving over critical information. Attackers or phishers pose as legitimate internet users. Phishers try to gain unauthorized access to a victim's accounts in order to steal sensitive or personal information and the victim's identity. The link will take the user to an impersonated website that asks for private information like usernames, passwords, or credit card details (Al-Momani and Gupta 2013).

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

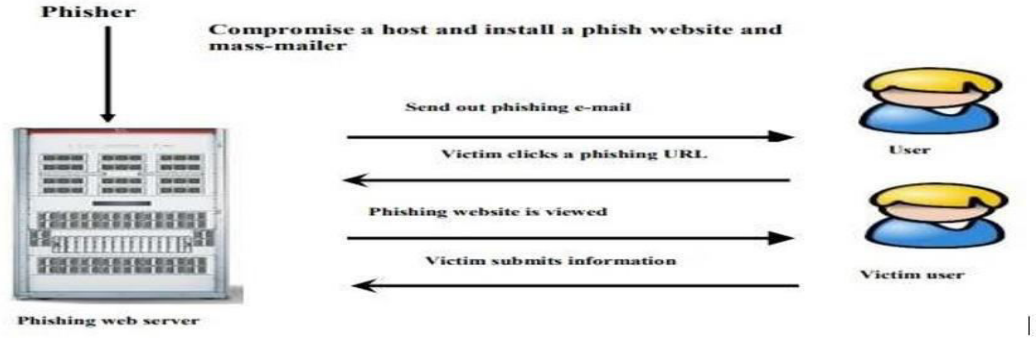


Figure 1: Phishing Lifecycle

The first method relies on social engineering schemes and involves sending fake emails that appear to be from legitimate businesses or bank accounts but actually lead the recipient to a fake website that requests confidential information like usernames, passwords, credit card numbers, and personal information.

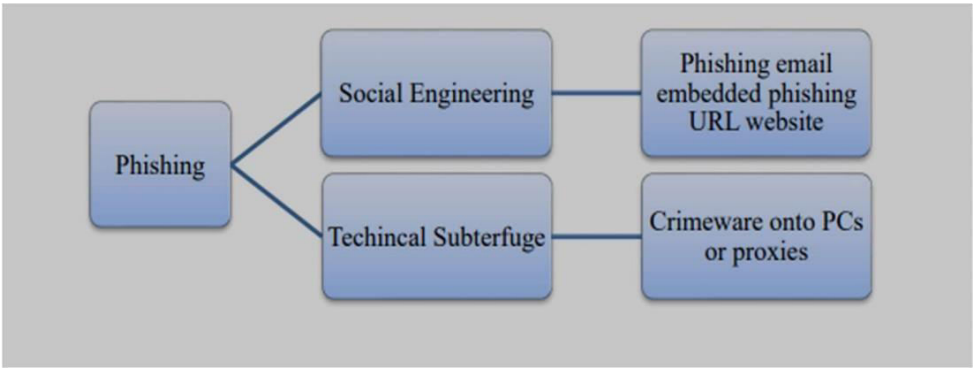


Figure 2: Types of Phishing E-mails

The malware-based phishing technique, however, relies on malicious codes or malware and technical schemes if users click on the embedded link or looks for security gaps in the receivers' devices to directly obtain their online account information. This method does not directly ask for details; instead, it relies on malicious codes or malware and technical schemes. The user may occasionally be misdirected by a phisher to a real website or one that is being watched by substitutes (Al-Momani, 2013) In 2012, an online report detailing an estimated $1.5 billion loss was published. The study blamed phishing attempts for the loss. Finding more effective phishing email detection techniques is necessary to contain the harm and lower the risk as a result of this enormous loss and threat, which is on the rise (Akinyelu, 2014).

2. LITERATURE SURVEY

There have been other surveys about phishing detection, as was already indicated. In order to start, this section summarises and cites the previous surveys. On the basis of 1-uniform resource locators (URLs), 2-websites, and 3-emails, the examined research have been categorised.

To categorise phished emails, Andronicus et al. employed a random forest machine learning classifier in [1]. They tried to reduce the number of features needed for classification while maximising accuracy. A highly accurate technique to content-based phishing detection is provided.

The authors of [2] suggested a model based on features that were retrieved from email headers and HTML bodies and were then categorised using feed forward neural networks. According to the findings,

the classification was accurate to 98.72%.A dataset of more than 7000 emails and a variety of attributes are employed in [3]. It is possible to reach 99.5% overall accuracy.

The goal of [4] by Gilchan Park et al. was to extract robust traits that could distinguish between genuine and phished emails. Between phishing emails and genuine emails, there are comparisons of sentence syntactic similarities as well as the distinction between the subjects and objects of target verbs.

The various phishing strategies are examined in [5], "Email Phishing: An Open Threat to Everyone," along with advice on how users might prevent falling victim to scammers.

In [6]. A methodology that combines natural language processing, machine learning, and image processing is proposed by Emilin Shyni et al. They employ a total of 61 characteristics. They had a success greater than 96% classification accuracy utilising several classifiers.

18 features are retrieved in [7] "Detection Phishing Emails Using Features Decisive Values," and the suggested algorithm categorises each email based on the presence of flags and the weighting of the features. Their findings demonstrate that if the most useful features are employed for classification, great accuracy can be attained from the 18 retrieved features. [8] The creators of "Phish-Detector" concentrate on the characteristics of Message-IDs and use n-gram analysis on the Message-IDs.

phishing attacks and their accompanying solutions are described in [9], providing insight into the characteristics that make phishing assaults easy to recognise. They compared 15 ways for detecting phishing assaults with another 15 techniques for identifying phishing websites. They also looked at several techniques from 2000 to 2016, separating phishing emails from real ones.

The researchers also provided a taxonomy or classification of these tactics and listed criteria that help identify between phishing and legitimate emails in [10], which also specified a few dataset sources and discussed a total of 18 potential solutions between the years 2000 and 2016.

They compared and assessed various anti-phishing tools that have been utilised in research and practise. By highlighting the assault vectors and communication methods that have hardly ever been discussed in literature, the researchers were able to pinpoint the gaps.

In this regard, Chiew et alsurvey's from 2018 [11] focused on the channels or vectors utilised in phishing assaults. The study explains the methods used in social engineering attacks as well as how they occur. Additionally, it highlights the potential for future attacks to be stronger due to the fusion of several strategies already in use.

A combination of the Naive Bayes classifier with the Apriori algorithm was proposed by Ishtiaq et al. [12] as an SMS spam categorization system. They combined association rule mining using the Bayesian and Apriori algorithms. Apriori collects the most frequently occurring terms that appeared together, and Bayesian then evaluates the likelihood that a word will appear both separately and in combination with other words in spam or ham communications.

Gomez et al. [12] examined how well Bayesian filtering techniques, which are used to prevent email spam, can be used to identify and thwart mobile spam. They preprocessed the communications using various tokenization techniques, picked out features, and evaluated their performance using several machine learning algorithms.

3. PROPOSED SYSTEM

According to the study conducted as part of the literature review, numerous researchers have carried out numerous studies and strategies to categorize or identify phishing emails, but many of them define spam and ham email filtering, while frequently some of them define proper emails filtering that are phished, many of them.
The users make use of blacklist allies, heuristics, and visual commonalities. The best outcomes were

achieved using this machine learning method as opposed to others.
Many users attempt to secure email filters, but end up using spam filters. As an example, consider the bag-of-words method, which extracts the highest occurring words from emails and uses them to classify. This method doesn't work for email filtering, but it works extremely well for spam filtering because phished emails contain specific features that are only used in phishing attacks. This method claims that spam filtering cannot properly handle emails that contain these features. The primary goal of this research is to increase the reliability of email filtering that determines whether an email has been phished or not. Use of supervised and unsupervised machine learning algorithms like random forest, logistic regression, Naive bayes, and support vector machines is the primary goal.

4. METHODOLOGIES

The methodologies used to determine whether E-mail messages are spam/ham (not spam)

- Data cleaning
- EDA (Exploratory data analysis)
- Data preprocessing
  - Lower casing
  - Tokenization
  - Removing special characters
  - Removing stop words and punctuations
  - Stemming
- Vectorization.
- Model building.

Data cleaning: Data cleaning is the process of correcting or deleting inaccurate, corrupted, improperly formatted, duplicate, or insufficient data from a dataset. There are numerous potentials for data duplication or labelling errors when merging multiple data sources.

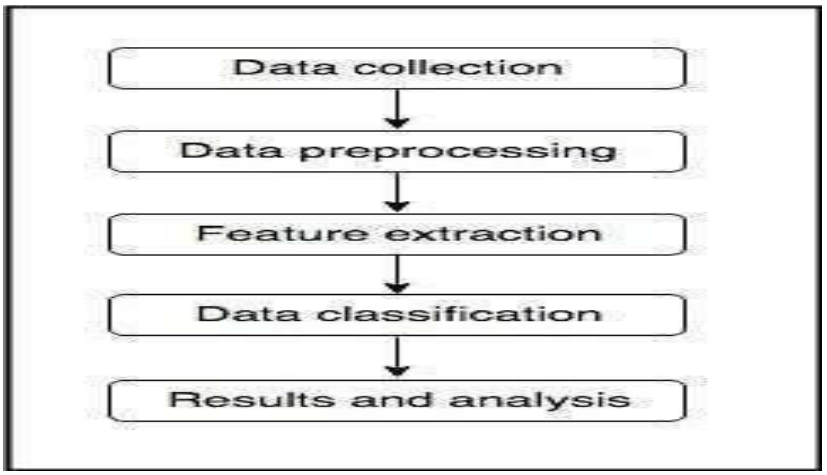


Figure 3: Process flow diagram words based

The quantity of action words in an email reveal whether the sender is anticipating a response from the recipient to do a particular action, such as clicking on a link, completing a form, or supplying certain information, etc. This is a permanent characteristic.
When the word "PayPal" appears, the sender frequently impersonates members of organizations that appear trustworthy. The word "PayPal" appearing in the email's links or "from" section may indicate

that the sender is affiliated with PayPal. It is a standard feature.
Word bank is present: This binary attribute implies that the letter is related to financial data. The sender would either be faking their affiliation with the banking

Model evaluation

We use a variety of matrices and classifications to evaluate the model using training datasets. 9.5339% of the training data are emails that have been phished. The training set consists of 29390 emails with the following 4 features: message content, subject content, message content type, and index. Only approximately 10% of the training data are emails that are phishing, therefore the labels are imbalanced. Each section is explained in the part that comes next.

A. As the value of the total dimensions is correctly separated in the accuracy fraction, the following can be

$$ACCURACY=(TP+TN)/(TP+TN+FP+FN)$$

B. Precision is incorporated into the critical metrics for the order concerns. To determine the assessed esteem, the precisely measured positive qualities are separated from the generic positive qualities. A greater precision rating is suggested by a lower bogus positive rate.

$$PRECISION=TRUEPOSITIVE/(TRUEPOSITVE+FALSE\ POSITIVE)$$

C. The F1-score measures how accurately a model fits a dataset. The F-score, which is classified as the consonant mean of the model's correctness and review, is a means of combining the exactness and review of the model.

$$F1SCORE=2*((PRECISION*RECALL)/(PRECISION+RECALL))$$

5. LIBRARIES AND PACKAGES INCLUDED

Matplotlib: For Python and its numerical extension NumPy, Matplotlib is a cross-platform package for graphical data visualization and charting. This makes it a strong open-source substitute for MATLAB. The APIs (Application Programming Interfaces) of matplotlib can also be used by developers to integrate plots into GUI programmes.

How to install Matplotlib: Set up Matplotlib the Python Package Index (PyPI) offers Matplotlib and its dependencies for download as a binary (pre-compiled) package. To install it, use the following command: Python with the pip option.

PANDAS: Working with data sets is made possible by the Python package Pandas. It offers tools for cleaning, examining, analysing, and modifying data.

Sklearn: Scikit-learn (Sklearn) is the most effective and reliable Python machine learning library. Using a consistent Python interface, it offers a variety of effective techniques for statistical modelling and machine learning, such as dimensionality reduction, clustering, and classification. NumPy, SciPy, and Matplotlib serve as the foundation for this library, which was primarily constructed in Python.

NumPy: In Python, NumPy is the core package for scientific computing. It is a Python package that offers a multidimensional array object, a number of derived objects (such masked arrays and matrices), and a selection of routines for quick operations on arrays, including mathematics, logical, shape manipulation, sorting, selection, I/O, discrete Fourier transformations, elementary linear algebra, elementary statistical operations, random simulation, and many other things.

Seaborn is a Python data visualization package built on the matplotlib framework. It offers a sophisticated user interface for creating visually appealing and useful statistics visuals.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

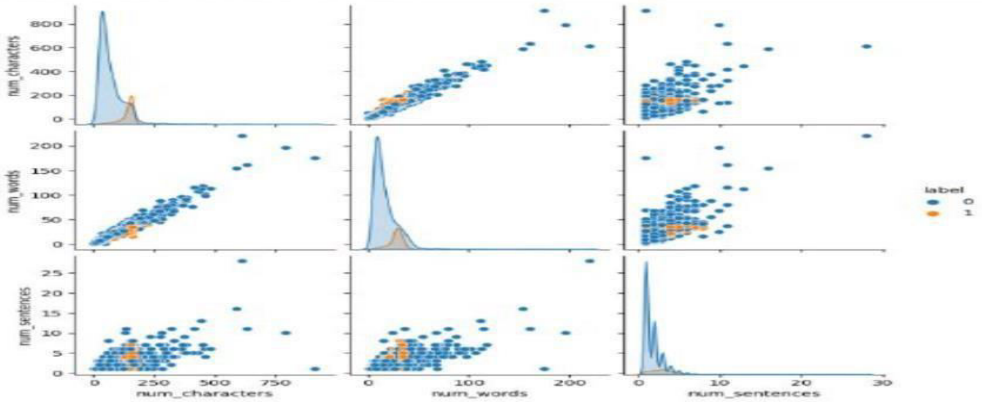


FIGURE 4: Seaborn

Non-functional requirements, as opposed to requiring specific behaviours, define criteria that can be used to assess how well a system performs. Functional requirements, which outline certain behaviour or functions, should be compared with this. Reliability, scalability, and affordability are examples of common non-functional requirements. The ilities of a system are another name for non- functionalities requirements. In addition to "requirements," "quality attributes," and "quality of service needs" are other terminology for non-functional requirements.

6. ALGORITHMS USED

Bayes Naïve: The Nave Bayes method is frequently employed as a training strategy for information repossession. The straightforward version was created to be used to extract the algorithm for analysing the relations file. The algorithm classifies the item based on each attribute's unique peek, and by using the Bayes law, one function can be identified. Each characteristic's presentation likelihood is concluded, and these chances are then **added up to produce a final** possibility. The probability is then determined for each class. Given the importance of the class variable, gullible Bayes procedures are a collection of controlled learning calculations based on applying Bayes' hypothesis with the "credulous" assumption of contingent freedom between each pair of provisions.

S.NO	Algorithm	Accuracy	Precision
1	KN	0.900387	1.000000
2	NB	0.979381	1.000000
3	RF	0.973888	1.000000
4	ETC	0.975822	0.982906
5	SVC	0.972921	0.974138
6	AdaBoost	0.961315	0.945455
7	LR	0.951644	0.940000
8	XGB	0.969052	0.934426
9	GBDT	0.952611	0.923810
10	BGC	0.958414	0.862595
11	DT	0.936170	0.846154

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

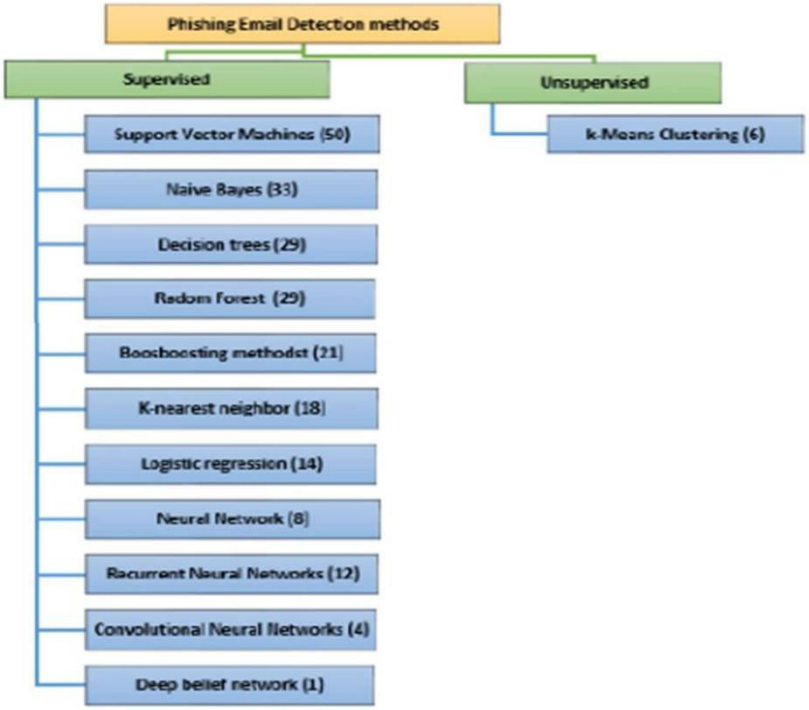
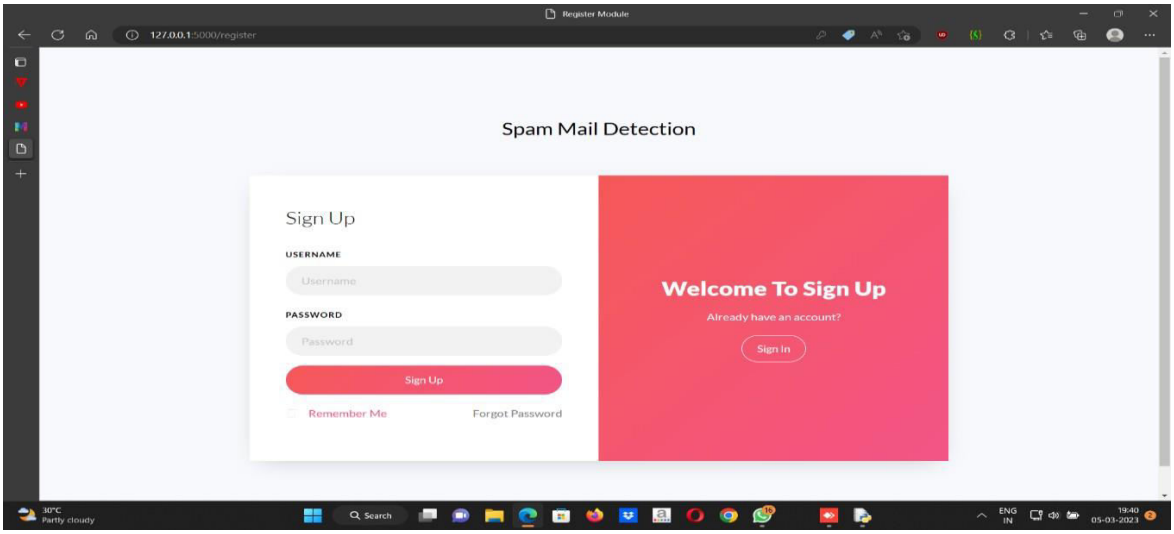


FIGURE 5:Algorithms used

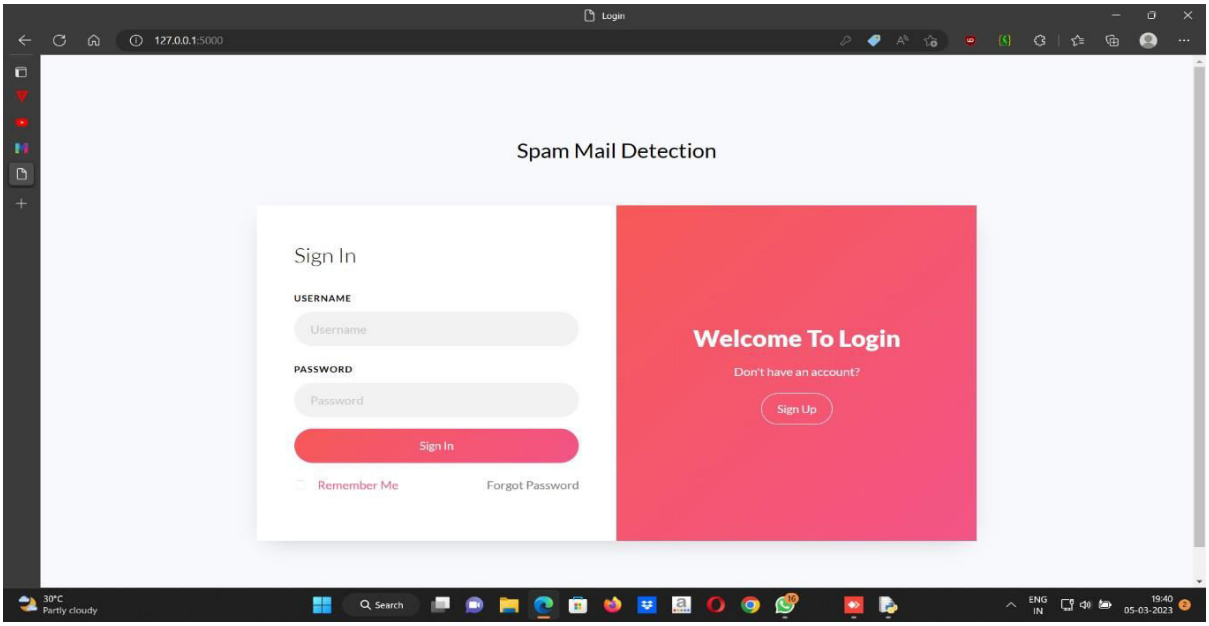
7. RESULTS

Registration Page



International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

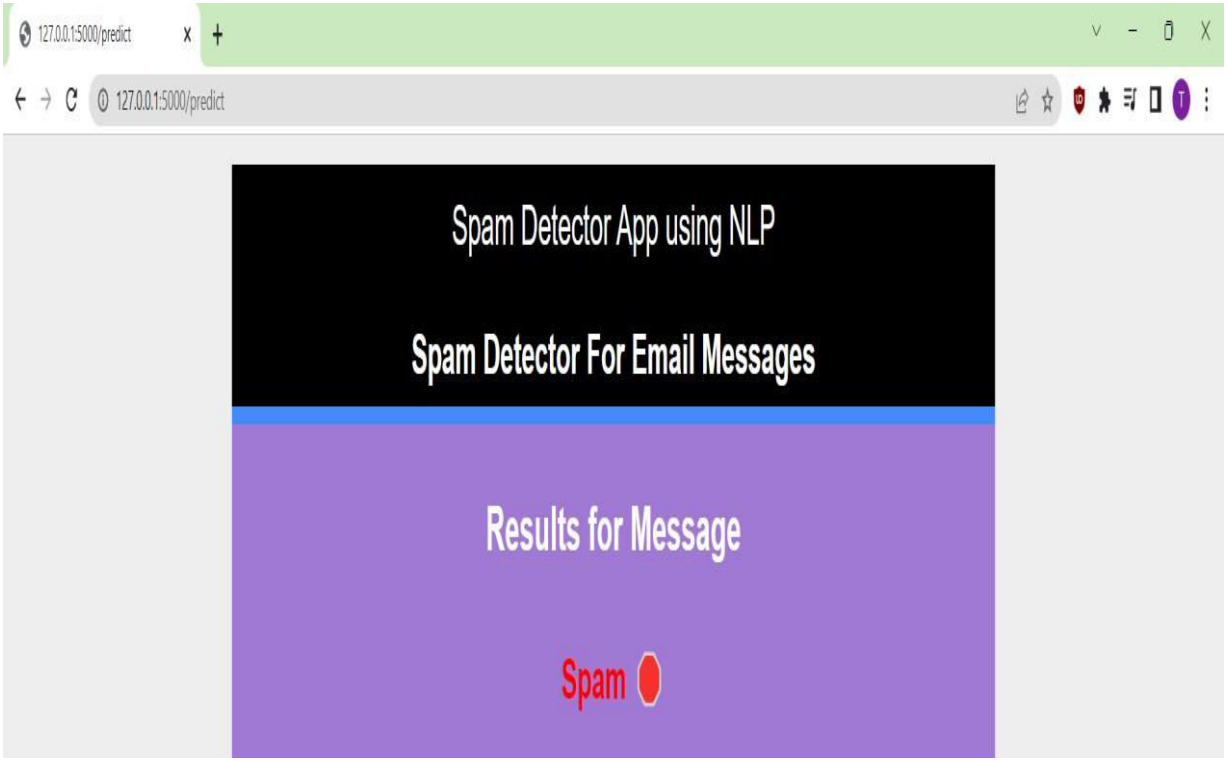
Login Page



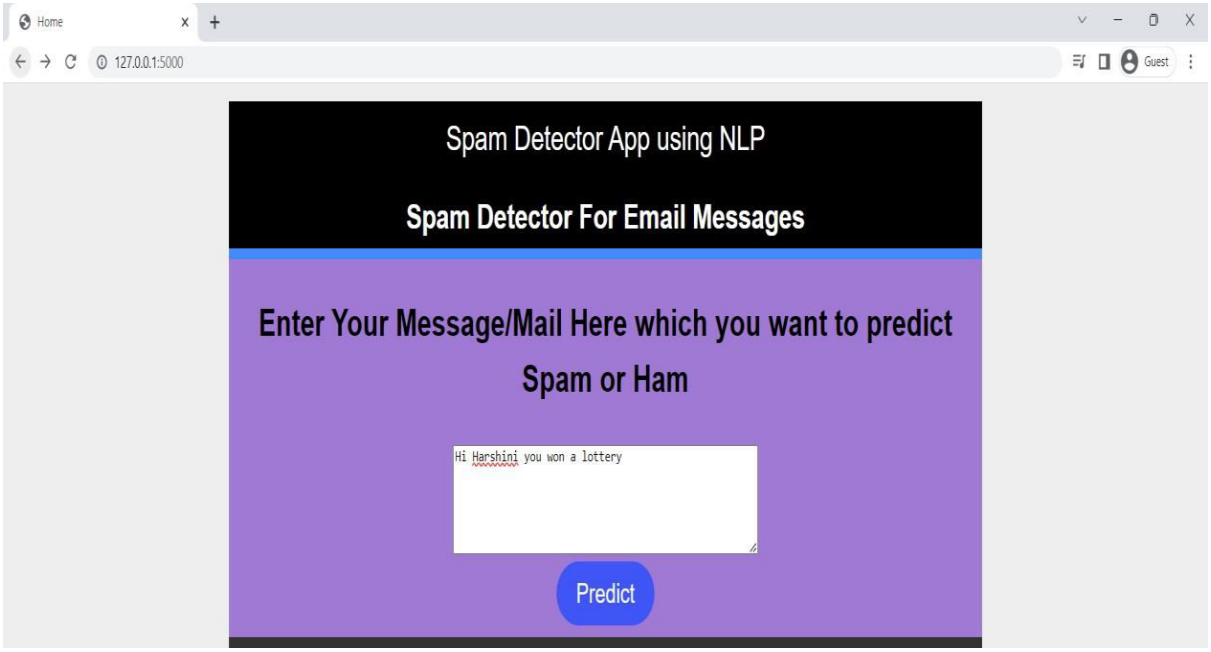
Home Page(Test 1)

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Result Page(prediction Page Test 1)

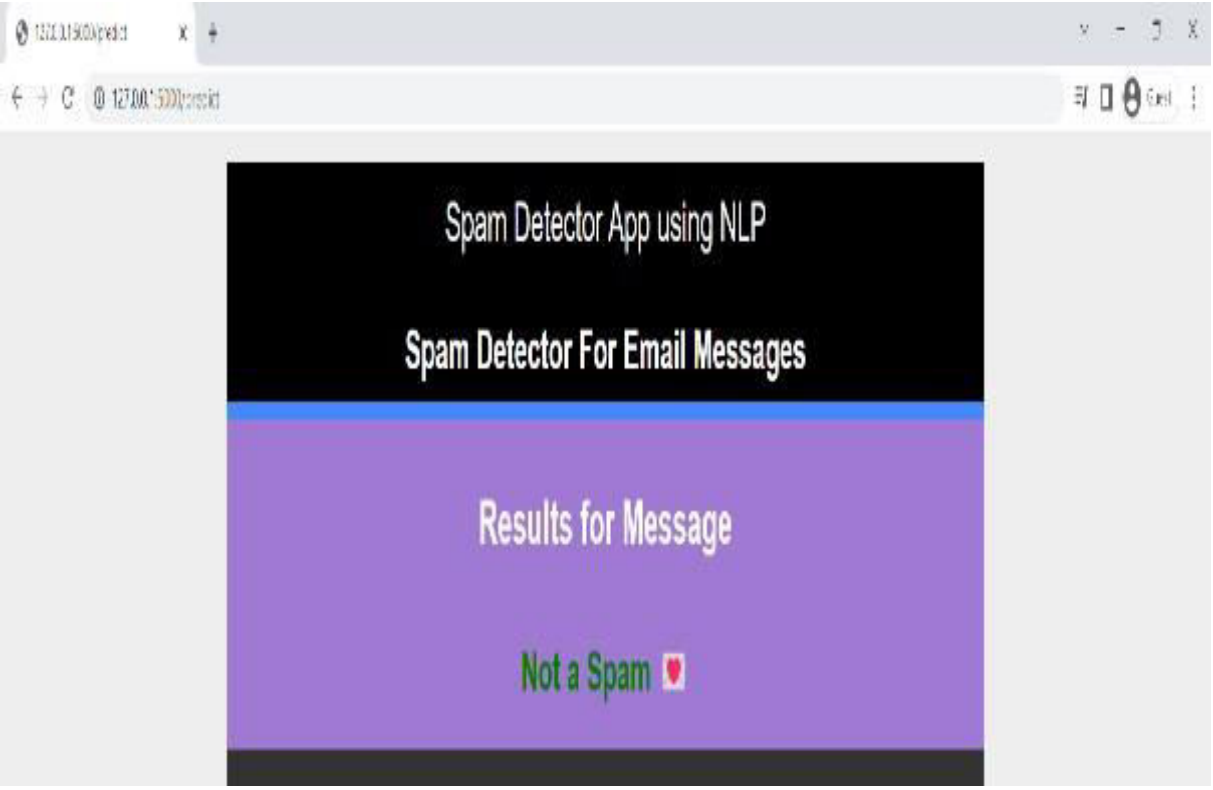


Home Page (Test 2)



International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Result Page (Prediction Page Test 2)



8. CONCLUSION

Phishing/Spoofing email detection is thought to be one of the most fascinating subjects in the field of cybersecurity at the moment. In order to evaluate the trend of phishing email detection, journal, conference, and workshop papers that were published between 2006 and 2022 were thoroughly examined using various methodologies. A thorough assessment of the literature was used to choose 100 publications. Many sorts of phishing email detection, for example, 'the domain name is misspelt,'\site email is badly written,' 'strange attachments or links'\sand 'phishing warning messages,' have been covered in our research. First, the basic details on the phishing ecosystem and useful phishing statistics are offered.

The taxonomy of phishing detection systems is then explained, and the datasets and features for detecting phishing emails as well as the detection algorithms and evaluation criteria are all covered. Finally, suggestions are offered to aid in the construction of phishing detection techniques so that compare-and-contrast schemes can be executed with ease.

This survey is distinctive in that it connects work to publicly accessible tools and resources. The study of the material that was presented showed that NLP approaches for detecting phishing emails had not received significant attention.

In light of the findings of the systematic literature review, it is clear that phishing email detection is the primary area of research, and the scientific community has made a concerted effort to address this issue in a number of widely used languages, including English. The Arab world is not an exception, but it has not been possible to extend those findings to the other cultures or situations, such as the developing non-English-speaking countries. The Arabic language is regarded as a Semitic tongue with a complex morphology. This supports our drive to eliminate language barriers by addressing the issue of Arabic-language phishing emails. Only a few articles on Arabic spam/phishing email detection using traditional machine learning are available. Studies on Arabic phishing email detection are insufficient due to a lack of resources for Arabic spam/phishing emails

and the slow progress made in dealing with Arabic NLP in general.

The main concept of this survey is to provide the most up-to-date tools and resources to the research community. The main goal is to inform the community of the benefits and drawbacks of each resource. It is clear that deep learning approaches have dramatically increased in popularity among researchers working on phishing email detection since 2019. The results that showed more research is needed before using contemporary deep learning techniques—such as long-short-term memory (LSTM) and CNN models—in phishing email detection studies. In this study area, the available resources and tools are insufficient. Therefore, additional research is urgently needed to evaluate deep learning approaches in the domain of phishing email detection.

After the study's thorough examination, several distinct observations - particularly in the field of machine learning-based proposal - were seen from the study's outcome, which was the cause for the great acceptance of supervised approaches. In addition, as mentioned, several algorithms, like NB and SVM, have important prerequisites. In addition, the bio-inspired Some researches have employed the computing (BIC) optimization technique, which has considerably improved classifier performance and decreased security concerns linked to costs of misclassification as well as user-dependent costs of misclassification.

As a whole, we can observe that single-algorithm anti-phishing systems are the most frequently used. Because of this, it is possible to analyse hybrid and multi-algorithm systems. Other than the URLs in the email body and the study that focuses on email header elements, the subject field and sender domain information must be carefully taken into account moving ahead. Also, the presentation of "Concept Drift" is a crucial area that could advance tools for spotting phishing assaults. Social honeypots and recommendation system algorithms are a novel idea that are employed for the identification of phishing that takes place between two malicious profiles as well as for the detection of similar phishing emails. This technique speeds up the process of phishing email detection. But there is a need for some creative solutions that may take into account all sides of a problem, as the most recent approach has not been as successful in dealing with the nature of phishing emails. Even governments of numerous major nations throughout the world, who have been criticised by numerous authorities, have been unable to come up with a successful method that has a lasting impact on this issue. Having said that, it has been observed recently that strengthening cybersecurity is given more importance.

REFERENCES

[1] Andronicus A. Akinyelu and Aderemi O. Adewumi. Classification of Phishing Email using Random forest Machine Learning Technique 2014.

[2] Noor Ghazi M. Jameel, Loay E. George. Detection of Phishing Emails using Feed Forward Neural Network, International Journal of Computer Applications 2013.

[3] Ian Fette, Norman Sadeh, Anthony Tomasi, Learning to Detect Phishing Emails, In Proceedings of the International World Wide Web Conference (WWW), 2006

[4] Gilchan Park, Julia M. Taylor, Using Syntactic Features for Phishing Detection 2015, https://arxiv.org/ftp/arxiv/papers/1506/1506.00037.pdf

[5] Gori Mohamed. J, M. Mohammed Mohideen, Mrs. Shahira Banu. Email Phishing - An open threat to everyone, International Journal of Scientific Research Publications, 2014

[6] C. Emilin Shyni, S. Sarju, S. Swaminathan A MultiClassifier Based Prediction Model for Phishing Emails Detection Using Topic Modelling, Named Entity Recognition and Image Processing, SciRes 2016

[7] Noor Ghazi M. Jamee , Loay E. George (2014), “Detection Phishing Emails Using Features Decisive Values”,257-259

[8] Rakesh M. Verma and Nirmala Rai. Phish-IDetector: Message-Id Based Automatic Phishing Detection, International Joint Conference on e-Business and Telecommunications 2015 .

[9] Basnet R., Mukkamala S., Sung A.H. (2008) Detection of Phishing Attacks: A Machine Learning Approach. In: Prasad B. (eds) Soft Computing Applications in Industry. Studies in Fuzziness and Soft Computing, vol 226. Springer, Berlin, Heidelberg

[10] Adwan Yasin and Adbelmunem, An intelligent classification model for phishing email