IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10424

Email Spam Detection and Prevention using Machine Learning

Jyoti Dake[1], Gunjan Memane[2], Prerana Katake[3], Samina Mulani[4]

Student, Department of Computer Engineering, Trinity College of Engineering & Research, Pune, India[1,2,3,4]

Abstract: As a means of contact for personal and professional use, emails are commonly used. Information shared that emails, such as banking information, credit reports, login details, etc., is often sensitive and confidential. This makes them useful for cyber criminals who are able to exploit the data for malicious purposes. Phishing is a technique that fraudsters use to acquire confidential data from individuals by claiming to be from proven sources. The sender will persuade you to provide personal information under bogus pretences in a phished email. Phishing website detection is an intelligent and efficient model focused on the use of data mining algorithms for classification or association. In order to identify the phishing website and the relationship that correlates them with each other, these algorithms were used to identify and characterize all rules and factors so that we detect them by their efficiency, accuracy, number of generated rules and speed. The proposed system integrates both classification and association algorithms, which optimize the system more effectively and faster than the current system. The error rate of the current system decreases by 30 percent by using these two algorithms with several protocols, so that the proposed system creates an effective way to detect the phishing website by using this approach. While there is no device that will detect the entire phishing website, it can create a more effective way to detect the phishing website using these methods. Spam emails can be only annoying but also dangerous to consumer. People using them for illegal and unethical conducts, phishing and fraud.

Keywords: Machine learning, Natural Language Processing (NLP), Feature extraction, Feature selection.

1. INTRODUCTION

Phishing is a lucrative type of fraud in which the criminal deceives receivers and obtains confidential information from them under false percentage . Phished emails may constrain to users to click on a link of a website or attachments where they are required to provide confidential information. The phisher sends phished emails to the thousands of users and out of which usually only a small percentage of recipients may fall into the trap but this can result in high profits for the sender. A huge phishing attack targeting millions of Gmail users hit Google in May 2017, in which the hacker gained access email histories of users. Through this information, the hackers were able to presents the emails as belonging to a known source and asked them to check the attached file. On clicking the link or attacked file, the users were asked to give permission for a fake app to manage users email account. With the ever increasing use of emails and growth of technologies, risk of losing valuable confidential information to fraudsters has also been increasing. This paper focuses on identifying the phished emails and senders with the help of machine learning algorithms.

In the proposed system, emails are classified as ham or spam. Machine Learning is a field of artificial intelligence in which the system is gives the ability to learn without being explicitly programmed. In our model, supervised machine learning algorithms are used for classification of emails. In our model using supervised machine learning algorithms email classification is done. These algorithms are a subset of machine learning algorithms which iteratively learn from large dataset. Electronic mail is a key revolution going down over standard communication systems because of its, fast, convenient, easy, and economical, to use nature. A main bottleneck in electronic communications is that the immense diffusion of unwanted, fraud, dangerous emails mentioned as spam emails. Machine learning (ML) researchers have developed various approaches to control this drawback. Inside this framework of machine learning, support vector machines (SVM) have ready an outsized half to the event of spam email filtering. supported Vector Machine, completely different theme are planned through text classification approaches. A essential drawback once victimization SVM is the choice of kernels as they freely affects the partition of emails within the standard area.

2. LITERATURE SURVEY

The systematic literature review provides us answers to research questions and queries, where as the general survey paper gives a broad idea about the email spam detection. SLR is performed by the guidelines provided and the selection of primary studies is performed . Objective :Research is to identify the extraction techniques, and present different existing models for email spam detection review and other available parameters to analyse these models. The population contains different keywords like Spam and Review, which are the keywords that are used for filter out the search records. In 1960 search procedure produced of initial studies. In this initial studies others are 165 are selected as

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10424

being relevant and 76 are selected as primary studies. The selection of primary studies is performed by three types of searches 1. primary search, 2.secondary search, 3.snowball tracking.
In 2007 ,Major research efforts in the domain of email spam detection review were started . Therefore, this all studies is based on a 12-year duration i.e.2007 to 2018. This research is executed a primary search by using different online research databases like Springer, Elsevier, IEEE, Science Direct, and ACM, conference proceedings, e-journals, and all review papers. In this study researchers used Google and Google Scholar to search the research databases.
1.Author: K. Krombholz, H. Hobel, M. Huber, and E. Weippl Title: Advanced Social Engineering Attacks, Journal of information security , applications. Proposed Approach: Attack scenarios of modern social engineering attacks on knowledge workers. Advantages: Introduced a comprehensive taxon-omy of attacks, classifying this attack by 1.attack channel, 2.operator, 3.different types of social engineering,4.specific attack scenarios. Disadvantages: It can just generate data but can't recommend any critical situations.
2.Author: E. Sorio, A. Bartoli, and E. Medvet 2013 Title: Detection of Hidden Fraudulent URLs within Trusted Sites :Lexical Features Proposed Approach: URL is fraudulent if the corresponding page is a deface-ment or a phishing attack (pages devoted to disseminating malware are beyond the scope of this work). The goal of the proposed method is to associate an input URL u with a Boolean value which indicates if u is an hidden fraudulent URL. Advantages: Proposed and evaluated an approach for the detection of hidden fraudulent URLs before actually fetching the corresponding page Disadvantages: Sometime it can predict false prediction.

3. PROPOSED SYSTEM

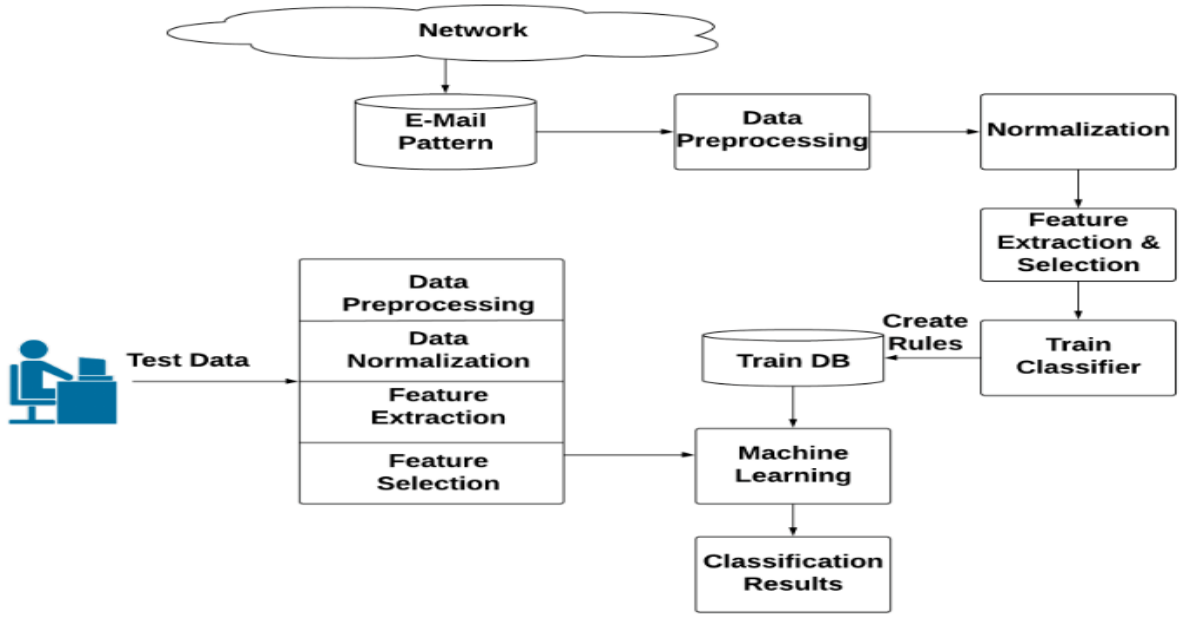

Figure 3.1: Proposed System Architecture

Proposed system
Description

- First system collect real time spam detection method was proposed using Machine Learning Classifier. Spam Mail Detection (SMD) which is classify email data into spam and ham emails.
- Email is one of the important type of Web Data Communication.
- The spam is undesirable information that a web user collect in form of mail or message. This junk is actually done by giving uncalled bulk message to indiscriminate set of receipts for publication or a commercial purpose.

Instruction/Training

- All collect data from the net or information server .Such as Methodical data as well as real time spam email information.
- Appeal data mining approaches like data bank or processing, data cleansing, asset, outlier detection and data transformation.
- Once complete the phase's data has saved into the data bank is called as background idea.
- It is used at the time of time discipline.

Examination/Testing

- First system collect real time and some real time mailing data and refer cross fold authenticate.
- All collected data has put into database or databank using object oriented connection architecture.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 4, April 2021

DOI 10.17148/IJARCCE.2021.10424

- In this testing we read all testing as well as trial data synchronously.
- Apply pre-processing on instruction and examination phase and then proceed with features of all.
- Train the system we have to use 2 different algorithms.

1. Machine Learning Algorithm
2. Generate Training Rules.

- Classify all test data and normal as well as spam based on calculator forest exam trial.
- Lastly predict the accuracy of overall system using different confusion matrixes and provide analysis precession with positive correct and Negative wrong of system.

4. ALGORITHM

4.1. Classification Algorithms

1) Fuzzy logic classifier with Decision Tree Mathematical Model
Input: database DB with each node having weight w and possible range for threshold T.
Output: Each instance with tree,disease probability
Step 1: Select max (w) instance from the database
Step 2: Select each max instance
Step 3: map w to the each conditional verifier from given threshold.
Step 4: set action label L to the each node.
Step 5: Generate tree with the label L
Step 6: Display tree.
Step 7: Save each tree node with disease into the databases.
End for
End for

2). Random Forest
Input: database DB with each node having weight w and possible range for threshold T.
Output: Each instance with tree ,disease probability
Step 1: Select max (w) instance from the database
Step 2: Select each max instance
Step 3: map w to the each conditional verifier from given threshold.
Step 4: set action label L to the each node.
Step 5: Generate tree with label L
Step 6: Display tree.

5. CONCLUSION

In this paper, we studied machine learning approaches and their application in the field of spam filtering. The samples of data sources and data collecting structures have led to a large increase in the data available for cyber security experts. To process such large volumes of data, scalable massive data processing solutions are needed. The present work on uses the Machine Learning algorithm which detects the spam emails but it gives accuracy around 70%. Our system will reduce the complexity along with that the accuracy increases and the spam emails will be detected successfully in less time. This review presents the how system works to detect the spamming a malicious contents from incoming emails using natural language processing and machine learning algorithms To detect the such entries system needs to analyse entire metadata of system and according to selected features it built training module. Different techniques are used in introduced to proposed review of supervised learning and detection analysis has done with respect to machine learning algorithms.

6. REFERENCES

[1] K. Krombholz, H. Hobel, M. Huber, and E. wiple“Advanced Social Engineering Attacks”, Journal of information security and applications 22 (2015) 113-122
[2] E. Sorio, A. Bartoli, and E. Medvt, “Detection of Hidden fallacious URLs in Trusted Sites using Lexical analysis Features”, International conference in 2013 on Availability and Security.
[3] M. Khonji, Y. Iraqi, and A. Jones, “Lexical URL analysis for discerning phishing and legitimate websites,” in Proceedings of the 8th Annual Collaboration, Spam Conference and Electronic messaging ser. CEAS ’11.
[4] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, “An empirical analysis of phishing blacklists,” http://ceas.cc/2009/papers/ceas2009-paper-32.pdf, 2009.
[5] Toolan, Fergus, and Joe Carthy. "Phishing detection and prevention using classifier ensembles." eCrime Researcher Summit in 2008. IEEE paper, 2009.