Email Detection and Prevention using Machine Learning

Prajakta Thorat[1], Sneha Hande[2], Harshada Tanpure[3]

1-3 Students, Department of Computer Engineering Jaihind College of Engineering, Kuran, Pune

---

Abstract - Nowadays, all the people are communicating official information through emails. Spam mails are the major issue on the internet. It is easy to send an email which contains spam message by the spammers. Spam fills our inbox with several irrelevant emails. Spammers can steal our sensitive information from our device like files, contact. Even we have the latest technology; it is challenging to detect spam emails. 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.

Key Words: Machine learning, Natural Language Processing (NLP), Feature extraction, Feature selection, classification.

1. INTRODUCTION

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 pretenses 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.

1.2. PROBLEM STATEMENT

For the purpose of classification, multiple packet features were extracted from all emails in a self-made dataset which consists of n number of phished emails and m number of ham emails. These features are fed into the classifiers and results noted.

Aim is to use the least number of features to develop a system which provides higher accuracy and study the variation of features and classify using various machine learning algorithms.

2. MODULES

i) Preprocessing –

The aim of pre-processing is an improvement of the image data that suppresses unwilling distortions or enhances some image features important for further processing, although geometric transformations of images (e.g. rotation, scaling, and translation) are classified among pre-processing methods here since similar techniques are used.

ii) Feature Extraction -

Feature extraction is a part of the dimensionality reduction process, in which, an initial set of the raw data is divided and analyze by SVM algorithm. So when you want to process it will be easier. The most important characteristic of these large data sets is that they have a large number of variables.

iii) Classification –

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

2.1. ALGORITHM

1. SVM

"Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot). Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line). It works really well with a clear margin of separation It is effective in high dimensional spaces. It is effective in cases where the number of dimensions is greater than the number of samples. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

2. LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feed forward neural networks, LSTM has feedback connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegment, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDSs (intrusion detection systems). A common LSTM unit is composed of a **cell**, an **input gate**, an **output gate** and a **forget gate**. The cell remembers values over arbitrary time intervals and the three *gates* regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

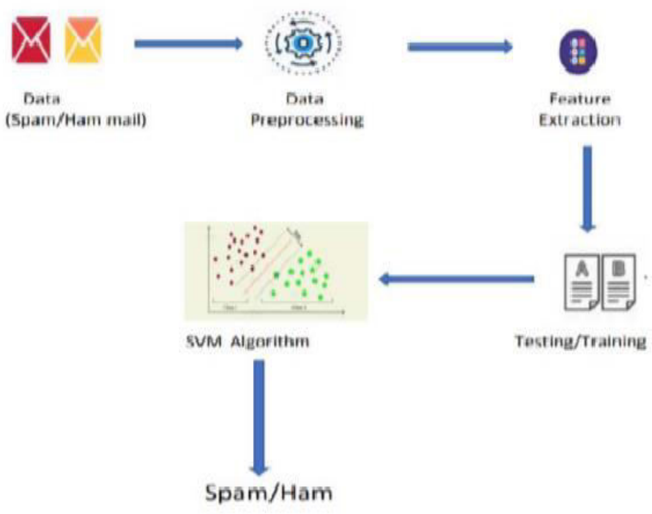
3. EXISTING SYSTEM

Machine learning based Spam E-Mail Detection had done by using Naive Bayes, J48 and SVM algorithms. It had less accuracy when compared to the proposed system.

4. PROPOSED SYSTEM

In the proposed system, detecting phished email can be described as a classification problem with two categories i.e. ham and phished. Whereas email which are meaningful with opposite nature are the ham. In our model, supervised machine learning algorithms used for classification. Supervised learning algorithms predict the nature of unknown data based on the known examples.

5. ARCHITECTURE



7. CONCLUSIONS

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 assistant how system works to detect the spamming a malicious contents from incoming emails using natural language processing and machine learning algorithms. To detect such entries system needs to analyze entire metadata of system and according to selected features it built training module. Different techniques help in introduced in proposed review for supervised learning and detection analysis has done with respect to machine learning algorithm.

ACKNOWLEDGEMENT

We would like to take this opportunity to thank our project guide Prof. Bangar.A.P. for giving us all the help and guidance we needed. We are really grateful to them for their kind support. Their valuable suggestions were very helpful. We are also grateful to thank, our other faculty members from the Computer Engineering Department, Jaihind College Of Engineering, Kuran for allowing us to perform our project.

REFERENCES

1. K. Krombholz, H. Hobel, M. Huber, and E. Weippl Advanced Social Engineering Attacks", Journal of information security and applications 22 (2015) 113-122
2. E. Sorio, A. Bartoli, and E. Medvet Detection of Hidden Fraudulent URLs within Trusted Sites using Lexical Features 2013

3. M. Khonji, Y. Iraqi, and A. Jones Lexical url analysis for discriminating phishing and legitimate websites 2011
4. S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang An empirical analysis of phishing blacklists.
5. F. Toolan and J. Carthy Phishing detection using classifier ensembles
6. Zhuang, L., Dunagan, J., Simon, D.R., Wang, H.J., Tygar, J.D., 2008. Characterizing Botnets from Email Spam Records, LEET'08 Proceedings of the 1st Usenix Workshop on LargeScale Exploits and Emergent Threats Article No. 2.
7. Enrico Blanzieri, Anton Bryl, 2008. A survey of learning based techniques of email spam filtering, Technical Report DIT-06-056.
8. Steve Webb, James Caverlee, CaltonPu, 2006. Introducing the Webb Spam Corpus: using Email spam to identify web spam automatically, CEAS.
9. Mishne, G., Carmel, D., Lempel, R., 2005. Blocking blog spam with language model disagreement.In Proc. 1st AIRWeb, Chiba, Japan.
10. Sculley, D., Gabriel M. Wachman, 2007. Relaxed online VSMs for spam filtering, SIGIR 2007 Proceedings.
11. Bing Zhou, Yiyu Yao, JigangLuo, 2010. A three-way decision approach to email spam filtering. Canadian Conference on AI, pp. 28–39.
12. MengjunXie, Heng Yin, Haining Wang, 2006. An effective defense against email spam laundering, CCS'06, October 30–November 3,Alexandria, Virginia, USA.