Survey Paper on Techniques of Email Fraud Detection System

Mr. Shrikant Salve[1] Mr. Sachin Rathod[2] Mr. Dheeraj Narkhede[3] Prof. Daivashala Deshmukh[4]
[1,2,3]Student [4]Professor
[1,2,3,4]Department of Computer Science & Engineering
[1,2,3,4]Maharashtra Institute of Technology, Aurangabad, India

Abstract— This survey paper categories a fraudulent email detection model using advanced feature choice. The paper of the literature study shows that the work of fraudulent emails detection requires the better or different choice of feature set; while the choice of classification method is of less importance. Fraudulent email can be evaluated using different state-of-the art algorithms. In comparison with related many reviews on fraud detection, this survey paper covers technical articles that proposes alternative data and solutions from related domains.

Keywords: Fraudulent emails, Spam emails

I. Introduction

Email is considered as a convenient way of written communication of this era. It is deemed to be an economical and steadfast method of communication. Email messages can be sent to a single receiver or broadcasted to groups. Main benefit of email message is that it can reach to a number of receivers simultaneously and instantly. These days, the majority of individuals even cannot expect the life exclusive of email. Now a day email becomes a widely used medium for communication of the people officially [2]. The continuous growth of the internet has also notably increased the number of email users. At the same time there is a noteworthy increase in spam emails rate. Emails can be classified in many groups, based on the purpose for which email is intended.

It can be classified into legitimate and illegitimate , spam and ham, suspicious and non-suspicious , fraudulent and normal, formal and informal which can further be classified as personal, family, friends, business, work, etc.

The broad category illegitimate email can be one that:

- When receiver is not interested.
- It is intentional for fraud purpose.
- The intention is to get crucial information from receiver.
- It might contain virus that harms receiver's computer.
- It might redirect receiver to illegitimate web site.

An email can be considered as illicit if it is not valuable for the receiver and for the society. Illegitimate emails might contain unwanted messages, phishing emails, threatening messages, or contain plans for some terrible events such as terrorist attack. Emails may have few features that, these can be sent anonymously without disclose the identity of the sender. In this paper we study the survey on fraudulent email detection by using various techniques, evaluating on well-known classification algorithms.

A fraudulent email is the one which is unwanted message; the receiver is not interested in. It is usually sent for deceiving purpose.

Some of the characteristics of such emails are as follows:

- Offering prize by greeting.
- It may contain financial terms, like money, share, and percent.
- Asks receiver to re email and contact as early as possible
- May talk about bad news like death of some person and gives greed to mail receiver.

II. Survey on Detection of E-mail Fraud

This paper shows study related to work and it is divided into two types. First study deals with the detection of the fraudulent emails, which are known as a kind of illicit emails, therefore, the survey shows for different illicit emails detection including spam, suspicious, phishing emails detection. Also another side of research regarding illicit emails is considered to be the authorship identification of anonymous emails.

A. Detection of Spam email

An email like spam is simply an email sent to address of any person which didn't ask for it to be sent. These emails are usually sent with different intentions, but advertisement and fraud are considered to be the major reasons. Spam email detection is mostly to be the classification task. It is usually said that there is no such technique which can provide complete solution against spam. Youn and McLeod [1] show a comparative study of various methods of classification for spam emails detection. In the comparative study, the authors presented Naive Bayes, SVM, J48, and neural networks classification techniques. The conclusion is that J48 classification is the most suitable technique for detection of spam email, because of the reasons the technique produced effective results.

Youn and McLeod another study based on spam filtering method. J48 algorithm used by author in order to formulate rules to generate concepts of the ontology.

B. Detection of Suspicious email

Another type of illicit emails are suspicious emails. Suspicious emails are those which consist of some material which is doubtful. For instance, an email may contain some text regarding some illicit activity; a threatening email. In the literature, the researchers also have contributed to this sensitive problem of suspicious email detection. Nizamani et al. [2] presented the suspicious email detection model based on enhanced feature selection. The authors used 'indicators' features in addition to the keywords for detection of suspicious email. Further, the authors emphasized on the use of the feature selection, in order to detect suspicious emails.

A Paper presented by Appavu et al. [5] used the association rule mining for detection of suspicious email. In the article [5], the authors added a specialized class of suspicious emails as an alert or the information using verb. An email is considered suspicious if in addition to keywords it contains future tenses to consider it as an alarm for future suspicious activity. It should be noted that in the articles [2], [5], the suspicious emails considered are the terrorism related emails which give some clue regarding future terrorist acts.

C. *Detection of Phishing email*

One of the types of illegitimate email is Phishing email, which is intentionally obtaining important email information from the receiver. Phishing problem is a complex problem, due to the fact that an attacker can easily make the duplicate. Phishing emails are the emails which are intentionally plan to acquire crucial information from the receiver. That information includes username, password, debit and credit card details, bank account information, etc.

The emails contain such content that the receivers immediately keep faith and turn to respond the email by clicking on the links provided in the email or send the important information in reply.

Chandrasekaran et al. [9], in their study presents detection of phishing email detection is a problem of classification and used style maker and structural features and applied SVM classification methods in order to detect phishing emails.

III. DETECTION OF THE FRAUDULENT EMAILS DIAGRAM

A. *Architectural Diagram*

The goal of the research is to classify and separate fraudulent mails from the normal email, with the intention that the receiver may not get confused from the fraudulent email.

The fraudulent email includes specific words, that, the receiver performs specific actions immediately which are harmful to make frauds. In this survey paper, we have considered detection of the fraudulent email as a classification problem for any classification problem needs a feature set and a classification algorithm. We have raw emails as input and in training each email is assigned a label/class fraud or normal. The fraudulent email detection process works according to the architecture, depicted in Fig. 1.

Fraudulent email detection architecture is comprised of six components, which works as an assembly of tasks which are described as follows.

1) Input Component:

The function of this module is to receive email contents as input. The emails in this module contain each part of the email, such as header and body.

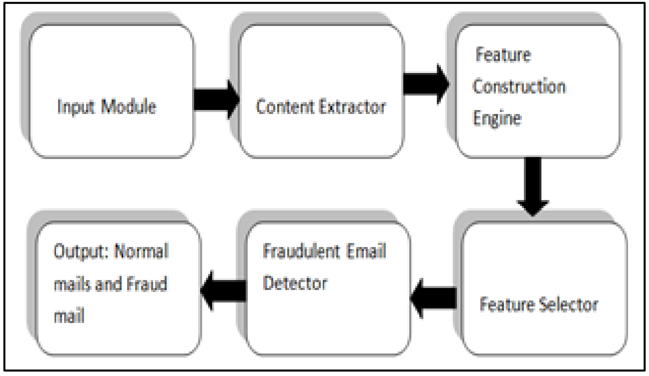
Input Module
Content Extractor
Feature Construction Engine
Feature Selector
Fraudulent Email Detector
Output: Normal mails and Fraud mail

Fig. 1: Detection of Fraudulent email Architecture.

2) Content extractor:

This module is implemented using Java code, which extracts the email content from raw email and saves it into the comma separated values (CSV) file format. Basically In this step architecture extracts the contents of the email, such as subject and body. The subject part of the email is extracted from the header of email, while the body is extracted as a whole. To recognize fraudulent mail the two parts like subject and body may extract from whole email because these two parts contain the text of the email

3) Feature construction engine:

Once the email content are available, feature construction engine builds up various feature-sets which are designed according to the experience and are found in various kinds of the fraudulent emails. Feature sets are divided into different categories, depending on type of fraud being considered in the email. Although, in the current work we only classify emails into fraud or normal but it is also possible to further classify fraud emails into different categories. Feature construction engine is implemented in Java.

4) Feature selector:

When different feature sets are available, not all features of worth considering for fraudulent email detection task. All features are assigned a weight using TF-IDF [10] scheme. Features are then separated as finance related and family related. The reason for separating the two types of features is to evaluate their usage in fraudulent emails. Analysis of fraud emails shows that the most of the fraud emails contain family and finance related terms. Analysis of the frequent terms in the emails is performed. These terms are then categorized into different sets. Classification models are then trained on these feature sets which give promising results.

Fraudulent email detector: This module applies the classification algorithms on features selected by the feature selector module. Various algorithms which are used for classification are applied using the machine learning tool WEKA which is an open source tool widely used by the research community in the area. The algorithms used for fraudulent email detection include: SVM, J48, Naive Bayes and CCM (cluster based classification model)

5) Output:

This module produces the results based on the features and classification algorithms used. The output is produced using the accuracy of fraud email detection, which is determined using 10-fold cross validation.

B. *Models used for Classification and Clustering*

We consider the task of fraudulent email detection as a classification task. In this section we discuss the classification algorithms we used for the detection of fraudulent emails.

1) J48

In the classification algorithms, decision tree method is one of the famous methods due to its simplification and inductive nature. J48 technique is WEKA's implementation of C4.5 [10], a well known decision tree algorithm.

2) SVM

Support Vector Machine (SVM) is widely used and considered as state-of-the-art classification method for text classification. It has an advantage over others that it can work well on high dimensional feature set. SVM has another advantage that it can transform non-linearly separable data to a new linearly separable data by using kernel trick.

3) Naive Baye's (NB)

NB is another well know algorithm used for classification, which uses Baye's theorem. It calculates the probabilities of the feature values for each of the classification category and uses these probabilities to predict the class of the unknown instances.

4) CCM (cluster based classification model)

CCM is a cluster based classification method, which performs the classification task by first grouping the data points based on obvious features. Once the groups of the instances are formed, SVM is applied to classify the instances in each of the cluster.

IV. Conclusion

Email fraud is a very serious issue that is not just annoying to the end-users, but also financially damaging and a security risk. This survey paper shows published paper related fraud detection studies. It defines the types and subtypes of fraud, the technical nature of data and the methods and techniques

References

[1] Youn S, McLeod D. Efficient spam email filtering using adaptive ontology. In: Fourth international conference on information technology, 2007, ITNG'07, IEEE; 2007. p. 249–54.Google Scholar

[2] S. Nizamani, N. Memon, U.K. Wiil, P. KarampelasModeling suspicious email detection using enhanced feature selection Int J Model Optim, 2 (4) (2012), pp. 371-377 CrossRefView Record in ScopusGoogle Scholar

[3] S. Nizamani, N. Memon, U.K. Wiil Detection of illegitimate emails using boosting algorithm, Counterterrorism and open source intelligence, Springer, Vienna (2011), pp. 249-264 CrossRefView Record in ScopusGoogle Scholar

[4] Sasaki M, Shinnou H. Spam detection using text clustering. In: 2005 International conference on cyberworlds, IEEE; 2005. p. 4 pp-316. Google Scholar

[5] Appavu, M. Pandian, R. Rajaram Association rule mining for suspicious email detection: a data mining approach Intelligence and security informatics, IEEE (2007), pp. 316-323 CrossRefView Record in ScopusGoogle Scholar

[6] McCallum A, Nigam K. A comparison of event models for Naive Bayes text classification. In: AAAI-98 workshop on learning for text categorization, vol. 752; 1998. p. 41–8. Google Scholar

[7] Joachims T. A statistical learning learning model of text classification for support vector machines. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM; 2001. p. 128–36. Google Scholar

[8] D.K. Renuka, T. HamsapriyaEmail classification for spam detection using word stemming Int J Comput Appl, 1 (2010), pp. 45-47 View Record in ScopusGoogle Scholar

[9] Chandrasekaran M, Narayanan K, Upadhyaya S. Phishing email detection based on structural properties. In: NYS cyber security conference; 2006. p. 1–7.Google Scholar

[10] S. Nizamani, N. MemonCEAI:CCM-based email authorship identification model Egypt Inf J, 14 (3) (2013), pp. 239-249, 10.1016/j.eij.2013.10.001 Elsevier ArticleDownload PDFView Record in ScopusGoogle Scholar