Egyptian Informatics Journal (2014) **15**, 169–174

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

ORIGINAL ARTICLE

Detection of fraudulent emails by employing advanced feature abundance

Sarwat Nizamani [a,b,*], **Nasrullah Memon** [a,c], **Mathies Glasdam** [a], **Dong Duong Nguyen** [a]

[a] *The Mærsk McKinney Møller Institute, University of Southern, Denmark, Campusvej 55, 5220 Odense, Denmark*
[b] *University of Sindh, Jamshoro, Pakistan*
[c] *Mehran University of Engineering and Technology, Jamshoro, Pakistan*

Received 24 October 2013; revised 7 July 2014; accepted 21 July 2014
Available online 13 August 2014

KEYWORDS

Classification;
CCM;
Feature set;
Fraudulent emails;
Spam emails

Abstract In this paper, we present a fraudulent email detection model using advanced feature choice. We extracted various kinds of features and compared the performance of each category of features with the others in terms of the fraudulent email detection rate. The different types of features are incorporated step by step. The detection of fraudulent email has been considered as a classification problem and it is evaluated using various state-of-the art algorithms and on CCM (Nizamani et al., 2011) [1] which is authors' previous cluster based classification model. The experiments have been performed on diverse feature sets and the different classification methods. The comparison of the results is also presented and the evaluation show that for the fraudulent email detection tasks, the feature set is more important regardless of classification method. The results of the study suggest that the task of fraudulent emails detection requires the better choice of feature set; while the choice of classification method is of less importance.

1. 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. An email message can reach to a number of receivers simultaneously and instantly. These days, the majority of individuals even cannot envisage the life exclusive of email. For these and countless other motives, email has also become a widely used medium for communication of the people having ill intentions [2].

[*] Corresponding author at: The Mærsk McKinney Møller Institute, University of Southern, Denmark, Campusvej 55, 5220 Odense, Denmark.
E-mail address: saniz@mmmi.sdu.dk (S. Nizamani).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

1110-8665 © 2014 Production and hosting by Elsevier B.V. on behalf of Faculty of Computers and Information, Cairo University.
http://dx.doi.org/10.1016/j.eij.2014.07.002

The rapid growth of the internet has also significantly increased the number of email users. At the same time there is a noteworthy increase in spam emails rate. A recent statistical report shows that the 70% of the email traffic during the second week of 2014 was spam[1]. As described earlier that fraudulent email detection is considered as classification problem, the research on email focuses on categorization of emails in different classes. Emails can be categorized in many groups, based on the purpose for which email is intended. It can be categorized as legitimate and illegitimate [3], spam and ham [4], suspicious and non-suspicious [2,5], fraudulent and normal, formal and informal which can further be classified as personal, family, friends, business, work, etc. [3].

The broad category illegitimate email can be the one that:

- Bothers the receiver means receiver is not interested.
- It is intended for deception purpose.
- It is intended to get crucial informat.ion from receiver.
- It may contain virus that harms receiver's computer.
- It may redirect receiver to illegitimate web site.

An email is considered illegitimate if it is not valuable for the receiver or for the society. Illegitimate emails may contain unwanted messages, phishing emails [6–8], threatening messages, or contain plans for some terrible events such as terrorist attack. Emails have other characteristics that these can be sent anonymously without revealing the identity of the sender.

In this paper we present the fraudulent email detection model by employing various features, evaluating on well known classification algorithms. A fraudulent email is the one which is unsolicited message; the receiver is not interested in. It is usually intended for deceiving purpose. Some of the characteristics of such emails are as follows:

- Greet by offering prize.
- Containing financial terms, like money, share, percent.
- Containing terms like advocate, and talking about some relation.
- Asks receiver to contact as soon as possible.
- May talk about death of some person and gives greed to receiver.

In this paper, we incorporated enhanced feature design for fraudulent email detection. The fraudulent emails are usually intended to cheat the receiver by tempting and showing helplessness to get the sympathies. Our dataset comprises of such emails which we consider deceptive and other emails that we consider normal emails. Considering the nature of emails we have used the features that can identify the emails of the kind, we specified. We conducted experiments using different feature sets and evaluated on various classification algorithms such as Naive Baye's (NB) [9], Support Vector Machine (SVM) [10], J48 [11] decision tree and CCM [1]. The experiments have been performed using well known open source machine learning tool WEKA [12].

The article is organized as follows: Section 2 discusses the related work, while Section 3 presents the fraudulent emails detection model. The experimental results are demonstrated in Section 4. Finally, Section 5 concludes the paper along with future directions.

2. Related work

Related work discussed in connection with the present study is divided into categories. This study deals with the detection of the fraudulent emails, which are known as a kind of illicit emails, therefore, the related work is presented for various illicit emails detection including spam emails detection, suspicious emails detection and phishing emails detection. Also another dimension of research regarding illicit emails is considered to be the authorship identification of anonymous emails. We also present some overview of the literature for email authorship identification.

2.1. Spam email detection

Spam emails are the illicit emails that a receiver is not interested in. The spam emails are unsolicited emails which are often sent in bulk. Spam emails are usually sent with different intentions, but advertisement and fraud are considered to be the major reasons. Spam email detection is often considered to be the classification task. It is believed that there is no such technique which can provide complete solution against spam. Youn and McLeod [13] presented a comparative study of various classification methods for spam emails detection. In the comparative study, the authors used Naive Bayes, SVM, J48, and neural networks classification techniques. The authors concluded that J48 classification is a suitable technique for the spam email detection task, because of the reasons the technique produced promising results.

In another study, Youn and McLeod [14] presented an ontology based spam filtering method. The authors used J48 algorithm in order to formulate rules to generate concepts of the ontology. The study by Renuka and Hamsapriya [15] adapted the use of word stemming instead of simply content based words for spam email detection. The authors showed that stemming based method is more efficient as compared to content based methods. It should be noted that Youn and McLeod [14] accentuated on the use of stemming based method, because the authors argued that the spammers use misspellings in order to deceive keyword based spam detection filters.

The most famous spam email detection filter "Spambayes" [16] used by Microsoft outlook as a plug-in uses Baye's theorem and uses keyword based approach for spam email detection.

2.2. Suspicious email detection

Suspicious emails are another category of illicit emails. Suspicious emails are those which contain some material which is doubtful. For instance, an email may contain some text regarding some illicit activity; a threatening email; or it may contain certain material which is worth analysis. Suspicious emails are deemed to be those which contain some clue regarding some illicit activities, which need to be further investigated by law enforcement agencies. There are some evidences regarding the exchange of suspicious emails before the events of 9/11 took place [23]. In the literature, the researchers also have

[1] https://www.securelist.com/en/analysis/204792327/Spam_report_January_2014.