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.

contributed to this sensitive problem of suspicious email detection. The study by Nizamani et al. [2] presented the suspicious email detection model based on enhanced feature selection. The authors employed the use of "indicators" features in addition to the keywords for suspicious email detection. Further, the authors emphasized on the use of the feature selection, in order to detect suspicious emails.

A study by Appavu et al. [5] applied the association rule mining for suspicious email detection task. 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.

2.3. Phishing email detection

Phishing emails are specialized class of illegitimate emails, which are intended to obtain useful information from the receiver of email.

Phishing problem is believed to be a security and privacy concern [6]. Phishing problem is considered to be the hard problem, due to the fact that an attacker can easily make the replicated website which may resemble to the legal bank of a user [7]. Phishing emails are the emails which are planned to acquire crucial information from the receiver. The crucial information includes username, password, credit card details, bank account information, etc. These emails resemble to the emails from trustworthy websites. The emails contain such a text that the receivers immediately turn to respond the email by clicking on the links provided in the email or send the crucial information in reply.

Chandrasekaran et al. [8], in their study consider phishing email detection as a classification problem and used style maker and structural features and applied SVM classification methods in order to detect phishing emails.

2.4. Email authorship identification

Email authorship identification is considered to be the task of identifying the most probable author of an email by analyzing the past emails of the suspected authors [17].

Li et al. [18] emphasized on the importance of writeprints in order to prevent cybercrimes. Authors argued that writeprints are as important as fingerprints are for identifying the criminals in real life.

The authors [19] presented a write-print based model for mining frequent patterns in the emails in order uniquely identify the authors of emails.

Nizamani and Memon [20] presented the model CEAI, which is CCM-based email authorship identification model. In the study, the authors employed traditional stylometric features along with their extended feature set and achieved promising results.

3. Detection of the fraudulent emails

The aim of the research is to separate fraudulent emails from the normal ones, with the intention that the receiver may not get affected from the fraudulent email in due course. The fraudulent emails often contain certain words, that, the receiver performs specific actions instantly which are harmful and result in frauds.

It should be noted that in this paper, we consider detection of the fraudulent email as a classification problem. For any classification problem one 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.

The architecture of the fraudulent email detection is comprised of six modules, which works as an assembly of tasks. The functions of each of the module are described as under:

Input module: This module is responsible for receiving email contents as raw input. The emails in this module contain each part of the email, such as header and body.
Content extractor: This module of the 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. The reason for extracting only the subject and body is that, these two parts contain the text of the email, which often describes the characteristics of the fraudulent email. 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.
Feature construction engine: Once the content of the emails 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.
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 [21] 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 [24] which is an open source tool widely used by the research community in the area. The algorithms used for fraudulent email detection include: SVM [10], J48 [11], Naive Bayes [9] and CCM (cluster based classification model) [1].[2]

[2] Details of each of the method are provided in later section.

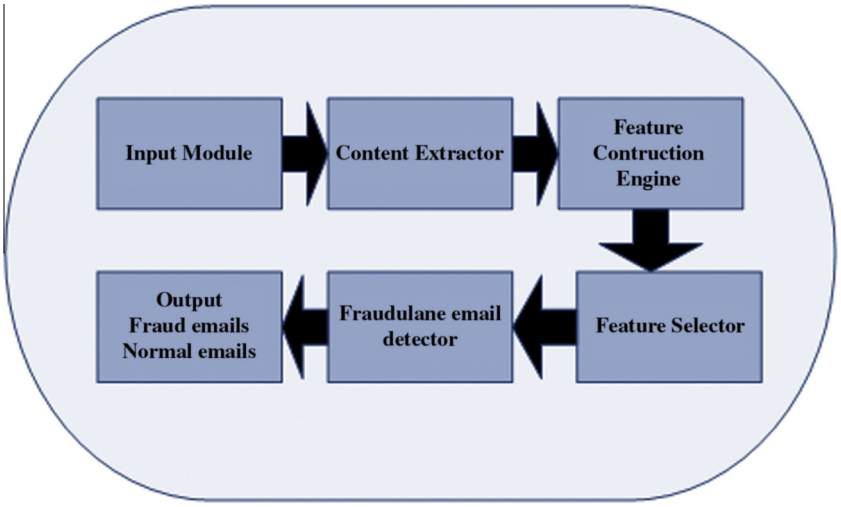

Figure 1 Fraudulent email detection architecture.

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.

The fraud email detection process works in flow, which is given Algorithm 1.

	Algorithm 1
Step1:	Input Email dataset E
Step2:	Extract email content
Step3:	Construct feature set from email content
Step4:	Make different feature selection
Step5:	For each set of features F_i
Step6:	For each set of classification algorithm C_j
Step7:	$Output_{i,j}$ = Fraud_detection (E, F_i, C_j)
Step8:	Display $Output_{i,j}$
Step9:	End loop at step5
Step10:	End loop at step6

Algorithm 1 describes the flow of the process, showing each of the steps of the fraud email detection process in sequence. Algorithm also depicts the process shown in the architecture.

First module of the architecture is the input module wyears old and I was diagnosed hich takes raw emails as an input, which contains the email content as well as the email header information. The content extractor module extracts the required parts of the email, i.e. the content which we use for the detection of the fraudulent email. Once, the content is available, the feature construction engine comes into the action. The 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.

3.1. Feature set

It should be noted that the fraudulent emails may be capricious; therefore, we build features-sets for different kinds of fraudulent emails. For example, some emails are intended to deceive the receiver by tempting them and showing helplessness for getting their sympathies in order to get receivers crucial information such as address and bank details and so on. Other emails may be more deceptive that look like that these are sent by the receiver's financial institute and require urgent action by the user and redirect them to some malicious website.

Initially features are extracted using the well-known scheme called TF-IDF [21], then based on heuristics, important features are extracted which have more capability to separate fraudulent emails from those of normal email. Afterward, family related and financial features have been separated.

An example fraudulent email containing family related terms, which are intended to deceive the receiver:

> "*I am Mrs. XYZ; I am a dying woman who has decided to WILL/donate what I have to you for the good work of humanity. I am 73 years old and I was diagnosed with cancer immediately after the death of my husband who has left me everything he worked for and my doctors told me I will not live longer than some weeks because of my health condition, that is why I decided to WILL/donate my money to you for the good work of humanity, and also to assist the less privilege, orphanages, widows and charitable organizations.*
>
> *I wish you all the best and may the good lord bless you abundantly, and please use the money well and always remember to extend the good work to others.*"

Lastly, the special features have been added that contain specific words in the subject of the email and contain hyperlinks in the body in order to redirect the receiver to a certain web-site. Table 1 shows the final-feature set that achieved the maximum accuracy.

Once the feature design is complete and the essential features have been chosen, the method for the detection of the fraudulent email has been applied. The fraudulent email detection process has been considered as classification problem, and experiments are performed using well known classification algorithms SVM, NB, J48 and CCM. Finally each email is

Table 1 Final feature set.

Donate	Customer
Buy	Pay
Account	Congratulation
Death	Please
Security	Deposit
User	Verify
$	£
Response	Attention
Dollar	Looking
Service	Valid
Urgent	Warning
Win	Won
Required	Offer
Risk	Money
Request	E-mail
Suspended	Transaction
Prize	Company

assigned a label fraudulent or normal. The experiments show the performance of the detection of fraudulent email by using various feature-sets.

3.2. Classification methods

We consider the task of fraudulent email detection as a classification task. Promising classification results can be achieved with the choice of representative features. In this section we discuss the classification algorithms we used for the detection of fraudulent emails.

3.2.1. 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 [11], a well known decision tree algorithm.

3.2.2. 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 [10].

3.2.3. Naive Baye's (NB)

NB [9] 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.

3.2.4. CCM (cluster based classification model)

CCM [1] 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.

4. Experimental results

For experiments, we used a dataset containing 8000 emails in total. Among 8000 emails half of the emails were fraudulent and half were normal. The fraudulent email dataset contained more than 2500 emails from Nigeria and are downloaded from the web site [22].

In order to conduct experiments, the emails are preprocessed and each email is represented as vector of features and an indicator of email type fraudulent or normal. A series of experiments have been performed using diverse feature sets and different classification methods. In the first set of experiments, we used features which are usually found in fraudulent emails and are intended to deceive the receiver by telling some family matters, and tempting them by offering some financial benefits in order to get some crucial information.

This basic set of features is comprised of only few features such as, father, mother, family, private, help and wife, husband. The results are shown in Fig. 2.

After the basic feature set comprised of family related terms, we performed experiments on finance related terms. All the four classification methods have been applied on this new feature set comprised of financial terms.

Finance related features are commonly found in almost all fraudulent emails, because the sender of such emails try to tempt the receiver by offering financial benefits in order to get his/her important information. The fraudulent emails that try to show helplessness to the receiver use family related terms as well as financial terms. Therefore, the accuracy of the fraudulent email detection increases significantly and reaches up to 89%, regardless of classification method. The results are illustrated in Fig. 2.

The next experiment has been performed using both of the features and frequency based features. Again, the accuracy of the task increases slightly.

The experiments conducted using family related, finance related and other frequency based features slightly increased the accuracy of task despite the consequences of classification method.

The last set of experiments has been performed using features which are commonly found in fraudulent emails but are rare in normal emails. In this new feature-set we extracted commonly found features from subject of the emails and a special feature which is found in rich text such as indicator of hyperlink in the body of email.

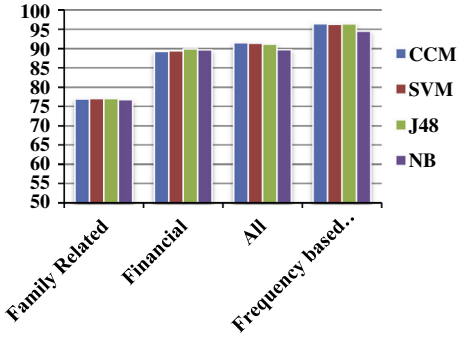

Figure 2 Experimental results.

The frequency based and intuitively chosen features increased the performance of the fraudulent email detection task. The final set of features attained maximum accuracy of the task because of intuitive features based on analysis of fraudulent emails.

The results of experiments of all feature types and different classification methods are illustrated in Fig. 2.

5. Conclusion and future work

In the paper we presented fraudulent email detection method, using advanced feature choice and classification techniques. We achieved the accuracy of fraudulent email detection as high as 96%. The research study also concludes that for the fraudulent email detection task, choice of efficient features affects the accuracy of the task. In the experiments we used various classification algorithms including SVM, NB, J48 and CCM. The experiments show that by including advanced features the accuracy of the detection of fraudulent detection task increases regardless of classification method because alike feature-set gives similar results for most of the classification methods. We conclude that the frequency based features attain high accuracy for the task of fraudulent email detection regardless of choice of classification method. In the current study, we have employed the features extracted from the content of the emails, by realizing the fact that often the fraudulent emails are characterized by content and we achieved the accuracy as high as 96%. However, we plan to employ header information of the email for the task of fraudulent email detection task.

References

[1] Nizamani S, Memon N, Wiil UK, Karampelas P. CCM: a text classification model by clustering. In: 2011 International conference on advances in social networks analysis and mining (ASONAM), IEEE; 2011. p. 461–7.

[2] Nizamani S, Memon N, Wiil UK, Karampelas P. Modeling suspicious email detection using enhanced feature selection. Int J Model Optim 2012;2(4):371–7.

[3] Nizamani S, Memon N, Wiil UK. Detection of illegitimate emails using boosting algorithm. Counterterrorism and open source intelligence. Vienna: Springer; 2011. p. 249–64.

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

[5] Appavu S, Pandian M, Rajaram R. Association rule mining for suspicious email detection: a data mining approach. In: Intelligence and security informatics. IEEE; 2007. p. 316–23.

[6] Dhamija R, Tygar JD. The battle against phishing: dynamic security skins. In: Proceedings of the 2005 symposium on usable privacy and security, ACM; 2005. p. 77–88.

[7] Fette I, Sadeh N, Tomasic A. Learning to detect phishing emails. In: Proceedings of the 16th international conference on World Wide Web, ACM; 2007. p. 649–56.

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

[9] 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.

[10] 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.

[11] Quinlan JR. C4. 5: programs for machine learning. Morgan kaufmann; 1993.

[12] Witten IH, Eibe F. Data mining: practical machine learning tools and techniques. Morgan Kaufmann; 2005.

[13] Youn S, McLeod D. A comparative study for email classification. Advances and innovations in systems, computing sciences and software engineering. Netherlands: Springer; 2007. p. 387–91.

[14] 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.

[15] Renuka DK, Hamsapriya T. Email classification for spam detection using word stemming. Int J Comput Appl 2010;1:45–7.

[16] Graham P. A plan for Spam. <http://www.paulgraham.com/spam.html> [accessed on 30.08.13].

[17] Zheng R, Li J, Chen H, Huang Z. A framework for authorship identification of online messages: writing-style features and classification techniques. J Am Soc Inf Sci Technol 2006;57(3):378–93.

[18] Li J, Zheng R, Chen H. From fingerprint to writeprint. Commun ACM 2006;49(4):76–82.

[19] Iqbal F, Hadjidj R, Fung B, Debbabi M. A novel approach of mining write-prints for authorship attribution in e-mail forensics. Digital Invest 2008;5(2008):S42–51.

[20] Nizamani S, Memon N. CEAI:CCM-based email authorship identification model. Egypt Inf J 2013;14(3):239–49. http://dx.doi.org/10.1016/j.eij.2013.10.001, Elsevier.

[21] Ramos J. Using TF-IDF to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning; 2003.

[22] Radev D. CLAIR collection of fraud email, ACL data and code repository, ADCR2008T001; 2008. <http://aclweb.org/aclwiki>.

[23] The 9/11 commission report; 2002. <http://www.gpo.gov/fdsys/pkg/GPO-911REPORT/pdf/GPO-911REPORT.pdf>.

[24] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 2009;11(1):10–8.