Soft Computing

A Hybrid Spam Detection Method Based on Unstructured datasets

--Manuscript Draft--

Manuscript Number:	
Full Title:	A Hybrid Spam Detection Method Based on Unstructured datasets
Article Type:	Original Research
Keywords:	Image spam; Text spam; Semantic networks; Classification; Subclass Discriminant Analysis; Feature Selection; Sparse Representation
Corresponding Author:	Marcello +441332591838 Trovati, Ph.D. University of Derby Derby, Derbyshire UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Derby
First Author:	Yeqin Shao
First Author Secondary Information:	
Order of Authors:	Yeqin Shao
	Marcello +441332591838 Trovati, Ph.D.
	Quan Shi
	Olga Angelopoulou
	Eleana Asimakopoulou
	Nik Bessis
Funding Information:	National Natural Science Foundation of China (CN) (61171132) — Yeqin Shao
	Natural Science Foundation of Jiangsu Province (CN) (BK2015022392) — Yeqin Shao
	Talent Project of Jiangsu Province of China (2014WLW029) — Yeqin Shao
Abstract:	The identification of non-genuine or malicious messages poses a variety of challenges due to the continuous changes in the techniques utilised by cyber-criminals. In this article, we propose a hybrid detection method based on a combination of image and text spam recognition techniques. In particular, the former is based on sparse representation based classification, which focuses on the global and local image features, and a dictionary learning technique to achieve a spam and a ham sub-dictionary. On the other hand, the textual analysis is based on semantic properties of documents to assess the level of maliciousness. More specifically, we are able to distinguish between meta-spam and real spam. Experimental results show the accuracy and potential of our approach.
Section/Category:	Methodologies & Application

Noname manuscript No.
(will be inserted by the editor)

A Hybrid Spam Detection Method Based on Unstructured datasets

Yeqin Shao · Marcello Trovati · Quan Shi · Olga Angelopoulou · Eleana Asimakopoulou · Nik Bessis

Received: date / Accepted: date

Abstract The identification of non-genuine or malicious messages poses a variety of challenges due to the continuous changes in the techniques utilised by cyber-criminals.
In this article, we propose a hybrid detection method based on a combination of image and text spam recognition techniques. In particular, the former is based on sparse representation based classification, which focuses on the global and local image features, and a dictionary learning technique to achieve a spam and a ham sub-dictionary. On the other hand, the textual analysis is based on semantic properties of documents to assess the level of maliciousness. More specifically, we are able to distinguish between *meta-spam* and real spam. Experimental results show the accuracy and potential of our approach.

Keywords Image spam · Text spam · Semantic networks · Classification · Subclass Discriminant Analysis · Feature Selection · Sparse Representation

1 Introduction

The ability of assessing malicious and non-genuine communication is crucial in all our activities, which are undeniably based on information sharing. For example, email communication has been an important means of communication in modern society due to its low cost and efficiency. There has been

Yeqin Shao

Tel.: +86 0513 85012470
E-mail: hnsyk@163.com

Marcello Trovati
Department of Computing and Mathematics
University of Derby, UK
Tel.: +44 1332 591838
E-mail: M.Trovati@derby.ac.uk

an increasing amount of research on spam, which includes unsolicited or malicious messages sent over the Internet, for the purposes of advertising, phishing, spreading malware, etc. [1]. However, there is a mounting argument regarding the concept of *non-genuine communication*. Spam emails are certainly part of non-genuine communications, where a user receives unwanted emails on a variety of topics. However, such type of communication can also be used to hide another message, such as the type of communication shared by terrorist cells after the 9/11 attack [2]. In a sense, the way terrorists attempted to share information is beyond the strict definition of spam. In fact, their communication was hidden into a non-genuine message. A full discussion into the differences between spam and non-genuineness goes beyond the scope of this article. As a consequence, we will use the terms "non-genuine message" and "spam" interchangeably, unless we wish to specify mutually exclusive features, and in such case, this would be clearly stated.

Non-genuine messages usually contain well-defined fragments of semantic networks based on specific keywords as well as on their usage. One of the most successful and efficient ways to detect spam messages focuses on meta-information contained in the message, e.g. sender's details, and title of the email. However, if we take a message without any such type of information, it is much harder to determine whether a message is genuine, especially when the message is carried out automatically. In fact, a sentence like "*buy a Rolex!*" would be identified as spam, in a similar fashion as "*asking you to buy a Rolex is very suspicious*" even though the latter does not appear to be spam.

Image spam usually transfers spam text such as advertisement, forged message, etc. onto an image, as depicted in Figure 3. The main properties related to an image include global and local features. The global features refer to the whole image, reflecting the overall characteristics of the image, which include colour, texture and shape. While the local features are based on a local image region, reflecting the details of an image. In this paper, such properties will be further investigated to provide a hybrid method to utilise in spam recognition. The other component is based on text analysis. Text based spam recognition often focuses on specific keywords, which might indicate the likelihood that a document, such as emails, is malicious [1]. However, this is not an easy task since, for example, an email discussing a fraudulent transaction could be potentially identified as malicious even though it may not be. Moreover, the current state-of-the-art methods and techniques often neglect the dynamics of the information extracted from textual sources. In fact, in a similar manner as in a variety of data and text mining tasks, information consists of "dynamic entities" and as such, its interpretation should consider its evolving nature [12].

The article is structured as follows: Section 2 gives an overview of the existing approaches, Sections 3.1 and 3.2 describe the methods and algorithms related to image and text based spam detection, respectively. In Section 4, the evaluation of the proposed methods is carried out, and finally, Section 5 concludes the paper and discusses future directions.

2 Related Work

In recent years, image spam has also become the effective communication channels of massive commercial advertisements and forged messages, which has led to the wide spread of spam and brings great challenges to the traditional text-based spam filtering.

Currently, image spam detection methods fall into three categories, as discussed in the rest of this section. The first category includes methods based on the approximation of the image properties in spam emails sent from the same source. Therefore, spam emails can be detected by clustering method based on the similarity of email images. Zhang et al. [3] and Chen et al. [4] have divided each image into three regions (i.e., text, foreground and background), and performed clustering based on the textual and visual features extracted from the image to detect the spams. Mehta et al. [5] represent the image with Gaussian mixture model, and cluster emails through Jensen-Shannon difference to identify spam. These methods have clustered spam emails based on the image features, thus the selection of image features will largely affect the clustering result. The second category includes methods based on text extracted from images by the Optical Character Recognition (OCR) technology, which can be used to detect spam emails characterized by such texts. Fumera et al. [6], Issac et al. [7], and Youn et al. [8] recognize texts in email images with OCR, and then identify spam emails with text filters. To avoid being detected, spammers usually disguise the images by text tilt, misspelling or misspelled, noise, etc. Byuu et al. [9] depict the image with the multiple-feature decision rules, and detect spam through MFOM learning method based on four major features. Dredze et al. [10] detect spam with maximum entropy and Bayes classifier based on the edge information and major colour range. Nhung et al. [11] employ the edge direction of characters in the image and supported vector machine to identify spams. To effectively detect spam, these methods are all needed to select discriminative features, and train strong classifiers. The key of image spam detection and classification lies in the discriminative features and the distinguishable classifier. To accurately detect image spam, this paper combines global and local features, and employs the subclass discriminant analysis to select the discriminative features, and then detects spam based on sparse representation based classification (SRC).

A well-known method to identify spam from textual sources, is based on Naive Bayes classifiers [1], which correlates the use of words in texts with spam and non-spam attributes. Bayesian inference subsequently calculate a probability that such text is indeed spam. Another successful approach is Enhanced TopicBased Space Model (eTVSM) [26] further improves the semantic evaluation by considering term interpretations. More specifically, the vector space defined by eTVSM is similar to the vector space of the TVSM, with the difference that document models are defined by interpretation vectors, which are created by analysing the formal procedure generated by the semantic relationships, such as word, word stem, term, interpretation and topic. To fully capture such

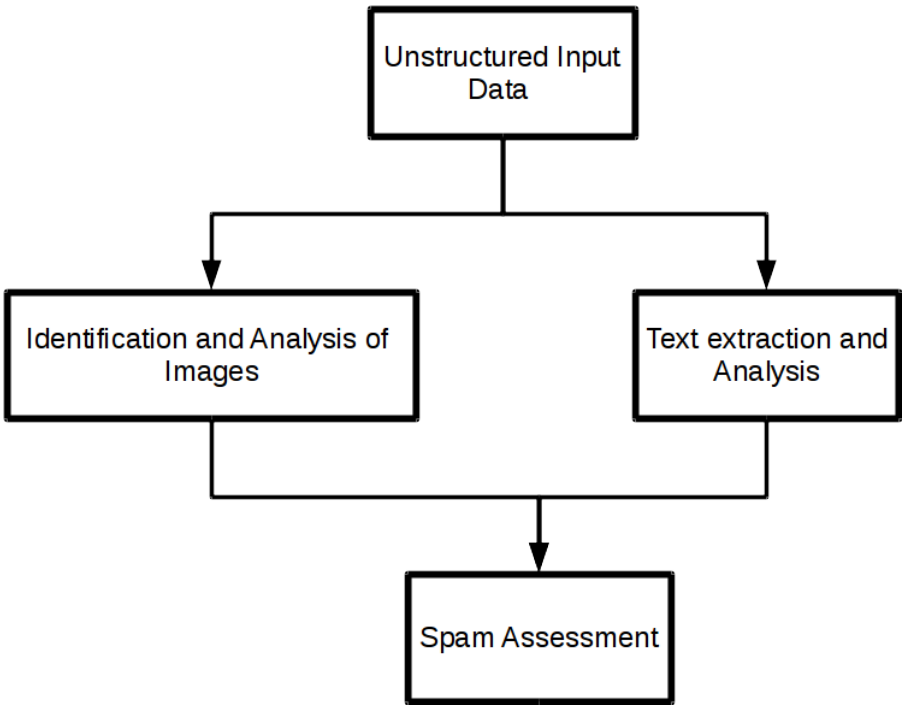

Fig. 1 The general architecture of the hybrid method proposed in this paper

semantic relationships, term relations are expressed in an ontology which operates with term, interpretation and topic concepts.

3 Method

In this section, we discuss the main components of our approach, i.e. image and text spam recognition, as well as their implementations. Figure 1 gives an overview of the main components of the approach introduced in this paper. In particular, Figures 2 and 4 depict the flow of the image and textual analysis, as discussed in Sections 3.1 and 3.2, respectively.

3.1 Image Spam Recognition

As discussed above, image features include global and local features. Among the former, the main features of the colour of an image include mean, variance, deviation, entropy and dispersion of image histogram; the texture features refer to the mean, variance, deviation, entropy and dispersion of LBP [13] image histogram; the shape features refer to mean, variance, deviation, entropy and dispersion of the gradient magnitude and gradient direction histogram of an

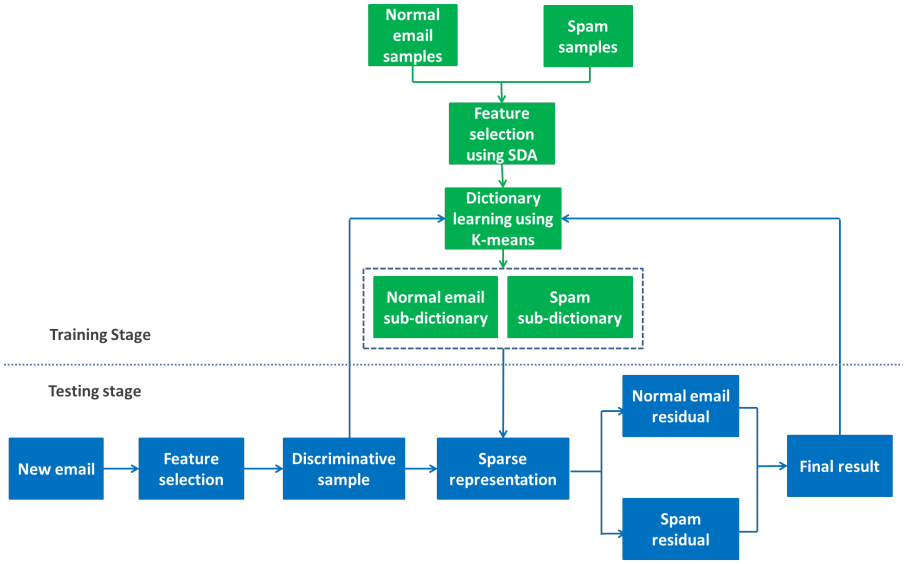

Fig. 2 Flowchart of spam Detection based on Sparse Representation based Classification

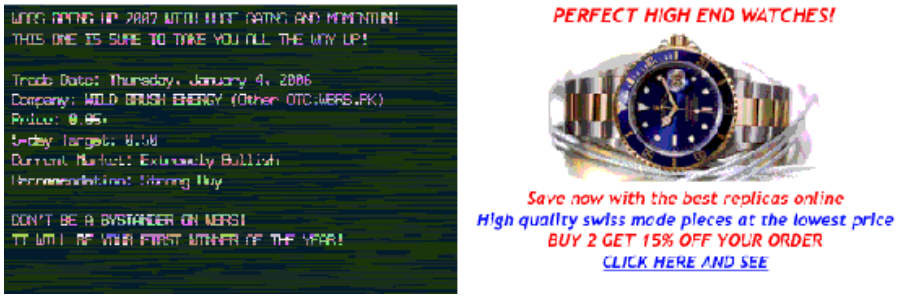

Fig. 3 Illustration of Image spam

image.

Among local features, scale-invariant feature transform (SIFT) can effectively address the disturbance in image spam, since SIFT is invariant to scale and rotation, and also robust to additive noise, affine distortion, and light changes. Due to the diversity and complexity of image spam, we take into account both global features and local features to accurately identify image spam.

3.1.1 Sparse Representation based Classification

Sparse representation reconstructs a new element with the linear combination of few dictionary elements, which are populated via dictionary learning technology. As opposed to other methods such as wavelet transform or Fourier transform, sparse representation does not require orthogonal or predefined dictionary elements. The process of dictionary learning is a task-driven proce-

dure and different tasks result in different dictionaries, whose elements need to be "representative". In particular, the aim of sparse representation is to distinguish multiple classes, suggesting that the elements in dictionary must be "discriminative".

Suppose $D \in \mathbb{R}^{n \times N}$ is a learned dictionary, which contains $n \times N-$dimension base elements. The sparse representation selects some elements in dictionary to represent a new element $x \in \mathbb{R}^n$. The sparse coding is as follows

$$c^* = \underset{c}{\operatorname{argmin}} ||x - Dc||_2^2 + \lambda ||c||_1, \tag{1}$$

where $c \in \mathbb{R}^N$ is the sparse code of x regarding the dictionary D, $|| \cdot ||_1$ is a $L1-$norm, λ is a coefficient used to control the sparsity, i.e. the number of non-zero elements of c^*. In particular, larger values of λ imply sparser values of c^*. In recent years, sparse representation based classification [3] has achieved promising result in face recognition. In the sparse representation based classification, to classify a new sample, we represent it with training samples from different classes simultaneously. The classification label is determined by the class with lowest representation error. Specifically, the training samples of the same class are first column-wisely combined into a sub-dictionary. Then all sub-dictionaries from different classes are further combined into a final global dictionary as follows

$$\begin{aligned} D &= [D_1, \ldots, D_i, \ldots, D_M] \\ D_i &= [d_{i,1}, d_{i,2}, \ldots, d_{i,N}] \end{aligned} \tag{2}$$

where D_i is the sub-dictionary of the $i-$th class, M is the total number of classes, $d_{i,j}$ is the $j-$th training samples from the $j-$training sample from the $i-$th class, $N = \sum_{i=1}^{M} N_i$ is the total number of training samples across all classes, where N_i is the number of training samples from the $i-$th class. To classify a new sample $x \in \mathbb{R}^i$, we apply (1) on the global dictionary D to get the global sparse code $c^* \in \mathbb{R}^N$, and then compute the representation residual r_i of each class, respectively. Therefore, we have

$$r_i = x - D_i c_i^*, \tag{3}$$

where c_i^* is part of the global sparse code corresponding to the sub-dictionary D_i. Subsequently, a new sample is classified into the class with the minimal norm of residual. Although the method is effective in face identification, it is not applicable to the massive spam detection for the two reasons. Firstly, due to the large number of spam, it is not feasible to include all training samples in the global dictionary, which will largely increase the computation of sparse representation. Secondly, the similarity of spam and normal emails entails the fact that the conventional hard classification method tends to generate the classification error, which cannot be corrected in later stage.

3.1.2 Improved Sparse Representation based Classification

In spam detection, different email classes can potentially contain similar training samples, which will decrease the overall classification performance. On the other hand, the process of discriminative sub-dictionary learning is based on elements in different sub-dictionaries, which are as different as possible. In this article, we combine feature selection and dictionary learning technology to learn discriminative sub-dictionaries. More specifically, we first select discriminative features with feature selection to increase the identification of different training samples. Subsequently, we adopt the dictionary learning technology to learn a compact sub-dictionary. The highly discriminative features can increase the distance of samples in different classes, and reduce the distance of samples in the same class, e.g., linear discriminant analysis (LDA) [15]. In particular, to eliminate useless and redundant features, LDA attempts to find a projection vector to maximize the Fisher discriminant criteria,

$$J(\omega) = \frac{\omega^T S_B \omega}{\omega^T S_W \omega} \tag{4}$$

where S_B and S_W represent the scatter matrix of between-class and within-class, respectively, that is

$$S_B = \sum_{i=1}^{N} (\mu_i - \mu)(\mu_i - \mu)^T \tag{5}$$

$$S_W = \sum_{i=1}^{N} \sum_{j \in C_i} (f_{i,j} - \mu_i)(f_{i,j} - \mu_i)^T \tag{6}$$

where N is the total number of classes,μ_i is the mean of samples in the $i-$th class, μ is the mean of all samples, C_i is the set of samples in the $i-$th class, f_{ij} is the feature of the $j-$th sample in the $i-$th class. Note that the eigenvector with the maximum eigenvalue of matrix $S_W^{-1} S_B$ is the optimal projection vector ω^*. For two-class spam classification, the samples are actually projected into one-dimension feature sub-space, which tends to hinder the spam identification. Therefore, we assume that each class i can be divided into L_i subclasses. With the subclass discriminant analysis [16], S_B can be re-defined as:

$$S_B = \sum_{i}^{N-1} \sum_{j}^{L_1} \sum_{k=i+1}^{N} \sum_{l}^{L_1} p_{i,j} p_{k,l} (\mu_{i,j} - \mu_{k,l})(\mu_{i,j} - \mu_{k,l})^T, \tag{7}$$

where $p_{i,j} = P_{i,j}/Q$ and $P_{i,j}$ represent the weight and the number of samples of the $j-$th sub-class in the $i-$th class, respectively, $\mu_{i,j}$ is the mean of samples of the $j-$th sub-class in $i-$th class, and L_i is the number of subclasses in the $i-$th class, which is determined by clustering on each class with affinity propagation

[17]. By considering Equation 7, we maximise the distance between different subclasses. Note that

$$\operatorname{rank} S_B \leq \min\left\{\sum_{i=1}^{N} L_i - 1, \operatorname{rank} S_W\right\}, \tag{8}$$

which addresses the issue of insufficient rank. The feature dimension in SDA is finally determined by the feature vectors with 95% variation. After feature selection, considering the efficiency, the large number of training samples cannot be directly used to construct the dictionary. To alleviate the storage space and computation cost, it is essential to learn a compact discriminative dictionary. Currently, most of the dictionary learning methods are used in reconstruction [18–20], where a dictionary can successfully represent a new sample. In this article, we use K-means clustering method to learn a compact sub-dictionary for each class as it aims to cluster training samples and select clustering centres as sub-dictionary elements to keep the discriminative ability of training samples and achieve better classification ability.
Once the sub-dictionary of each class is learned, the dictionary elements are normalized column-wise, and then combined as (2) to achieve a global dictionary for classification.

3.1.3 Soft Classification based on Logistic Regression

Since the potential similarity between spam and genuine emails, we perform soft classification based on Logistic Regression to avoid the error of hard classification. More specifically, after sparse representation based classification, the residuals r_i of different classes are combined into a residual vector R_i to establish a residual space. Subsequently, the probability of each email being spam is assessed via Logistic Regression [21]. The Logistic Regression model can be obtained as follow:

$$\Phi(\alpha, \rho) = \sum_{i=1}^{K} \log\left(1 + \exp\left(-z_j(\alpha^T R_i + \rho)\right)\right) + \phi||\alpha||_2 \tag{9}$$

where α and ρ are the coefficients of Logistic Regression, R_i is the combinational residual vector of the $i-$th sample, z_i is the label of the $i-$th sample, K is the total number of training samples, ϕ is a regulation coefficient, $||\cdot||_2$ is a $L2-$norm, which is used to avoid over fitting. With the optimal coefficient α^* and ρ^*, Logistic Regression can predict the probability of each sample

$$h(\mathbf{y}) = \frac{1}{1 + \exp\left(-(\alpha^{*T}\mathbf{R_y} + \rho^*)\right)} \tag{10}$$

where $\mathbf{R_y}$ is combinational residual vector of a test sample $\mathbf{y}$. The label can be determined by comparing the probability $h(\mathbf{y})$ and a threshold Th. We, therefore propose Algorithm 1.

To keep our method adaptive to the change of spam, we need to update the existing dictionary with new samples. Based on a set of labelled emails,

Algorithm 1 Image Spam Identification

1: Let I be an image email
2: Draw samples on I
3: Extract the discriminative feature vector f_k selected by SDA at each sample
4: Sparse represent each f_k with the learned global dictionary D (Eqn. 1)
5: Compute the representation residual r_i for each sub-dictionary D_i (Eqn. 3)
6: Get the probability $h(y)$ in the residual space with the combinational residual vector (Eqn. 10)
7: Determine the label: if $h(y) > Th$, it is a spam; otherwise, it is a normal email.
8: **return** Label of the image email.

we achieve the original sub-dictionaries by $K-$means. After classification, a new email can be taken as a sample of the corresponding class, and used to update the corresponding sub-dictionary. In this way, the dictionary contains the latest email information, and the proposed method can achieve accurate classification.

3.2 Textual Identification of Non-Genuine Messages

The second part of this article focuses on an efficient and accurate method to identify and assess spam embedded in texts. As discussed in [12], the use of network theory can be very beneficial in terms of power of abstraction and generalisation, efficiency, and scalability.
In this section, we discuss how networks, and semantic networks more specifically, can be utilised to identify and assess the level of maliciousness of a message.

3.2.1 Textual Analysis Details

Figure 4 depicts the main architecture of the approach discussed in this section.
Let $G = G(V, E)$ be a network, such that $V = \{v_i : 1 \leq i \leq n\}$ is the *vertex set* and $E = \{e_{i,j}\}_{i \neq j=1}^{n}$ is the *edge set*. Suppose that G can be embedded onto two different hyperplanes H_1 and H_2, such that the former is the *semantic layer*, and the latter is the *non-genuine layer*, which contains the concepts related to spam messages and their mutual relationships.
More specifically, $H_1 = H_1(V_{H_1}, E_{H_1})$ where $e_{i,j} \in E_{H_1}$ if and only if $v_i, v_j \in V_{H_1}$ are linked by a semantic relationship, such as synonymy, and $H_2 = H_2(V_{H_2}, E_{H_2})$ where $e_{i,j} \in E_{H_2}$ if and only if $v_i, v_j \in V_{H_2}$ are connected by a non-genuine relationship.
In particular, H_2 contains sub-networks, not necessarily connected, describing concepts which are likely to be part of a non-genuine message, as well as their corresponding synonyms. However, only concepts do not provide enough information to assess whether a message is indeed of malicious nature. On the other hand, specific collections of such concepts linked by mutual relationships, lead to their full and successful assessment.

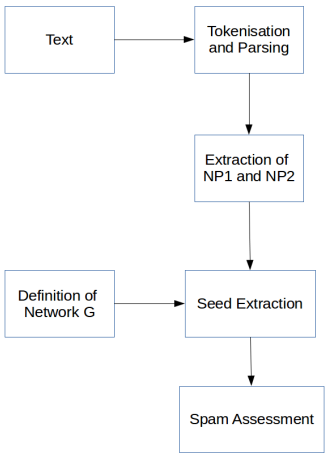

Fig. 4 The flow of the method discussed in Section 3.2.1

Consider, for example "*pornography is illegal in certain countries*" opposed to "*Buy cheap pornography!*". The former, unlike the latter, discusses an issue regarding the "pornography" concept, which is perfectly genuine. However, the term "pornography" can erroneously activate keyword-based semantic rules, suggesting that this message is spam. Therefore, to optimise the classification process, it is crucial to have a deeper understanding of the meaning of such sentences.
Loosely speaking, this issue can be formulated as the task of distinguishing spam messages from non-spam (more generally, *meta-spam*, i.e. messages, or texts *about* spam). In this paper, we make the assumption that if a fragment of a text, likely to be identified as non-genuine, refers to *direct* communication, i.e. between quotation marks, then it is regarded as non-spam. In particular, we define Algorithm 2.

The aim of the algorithm above is to be able is to incrementally update the network by considering newly available textual datasets. Each time a new text, or document is included, the steps above are carried out, which update the network by merging it with new nodes and edges.
The process of merging the X_i's into X is carried out recursively, so that identical nodes and edges are merged, as there is no loss of information. In fact, the relation `rel` between `NP1` and `NP2` has general semantic connotations, and so, there is no need to diversify them.

Algorithm 2 Text Analysis

Let $T = \{T_1, \dots T_k\}$ be a, potentially large, unstructured dataset, consisting of k texts.
$i = 0$
while $i \leq k$ **do**
    Parse T_i, and carry out anaphora resolution.
    Extract all triples `<NP1, rel, NP2>`, where `NP1` and `NP2` are the noun phrases connected by `rel`, which is a combination of verbs and specific keywords.
    A network $X_i = X_i(V_{X_i}, E_{X_i})$ is defined by all the couples of nouns/keywords within `NP1` and `NP2`
    Merge X_i into a weighted network $X = X(V_X, E_X)$
    For every edge $e_{i,j} \in E_X$, let its weight $w_{e_{i,j}} = \dfrac{|e_{i,j}|}{\sum_{e_{x,y} \in E_X} |e_{x,y}|}$
    $i = i + 1$
end while
return X

3.2.2 Seed Extraction

We define a *seed* as a sub-network $S \subset G$, where G is the network defined in Section 3.2.1. The level of non-genuineness of S is then measured depending on the number of components of $S \cap H_2$, i.e. the components of S that can be of malicious nature.
All the steps of this part are define Algorithm 3.

Algorithm 3 Seed Extraction

Let X be the network extracted from Algorithm 2
Let $X_S = X \setminus \bigcup_{i=1}^{N} s_i$ be the network created by removing the set of seeds S from X
loop
    Choose the seed set $S = \{s_i\}_{i=1}^{N}$ such that
Require:
    s_i are maximal
    $s_i \cap s_j$ is minimised for $i \neq j = 1, \dots N$
    for X_S **do**
        Minimise size(X_S)
        Either minimise $\sum_{e_{i,j} \in E_{X_S}} w_{e_{i,j}}$
        Or $\sum_{e_{i,j} \in E_{X_S}} w_{e_{i,j}} < T$, where T is the *weight threshold*
    end for
end loop
Let M be the non-genuineness parameter

$$M = \frac{\left|\left(\bigcup_{i=1}^{N} s_i\right) \cap H_2\right|}{\left|\bigcup_{i=1}^{N} s_i\right|} \left(1 - \frac{|X_S|}{|X|}\right).$$

return M

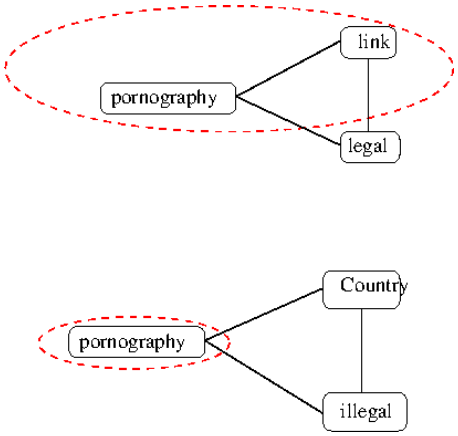

Fig. 5 Example of seed extraction as discussed in Section 3.2.2. The dotted lines encircles nodes that belong to H_2.

Clearly, condition 10 in Algorithm 3 provides a "soft" requirement compared to condition 9.

Note that $0 \geq M \leq 1$, so that if $M = 0$ the network does not contain *any* seed which may indicate a (totally) genuine message, whereas $M = 1$ would be the opposite, i.e. a completely malicious message.
Consider, for example the following two sentences

1. *Follow this link for legal pornography!*
2. *Pornography is illegal in this country.*

We can easily see that the first sentence is certainly more likely to be of malicious nature than the second one. In fact, according to the keywords described in Section 4, the seed extraction would identify *pornography*, *link* are nodes of H_2, whereas *illegal*, *legal*, and *country* belong to V_{H_1}. As depicted in Figure 5, it is clear that the first sentence will be identified as more malicious than the second one.

Algorithm 4 General Algorithm

```
Let F be a file, and assume it has either text, one or more images, or a combination of both
if F contains image(s) file then
    Start Algorithm 1
end if
if F contains a text file then
    Start Algorithms 2 and 3
end if
if F contains both image(s) and text file then
    Start Algorithms 1, 2 and 3
end if
return Outputs from above algorithms
```

Table 1 Illustration of classification results

Classification Results \ Actual Results	spam Email	Legitimate Email
spam Email	a	b
Legitimate Email	c	d

4 Evaluation

In this section we discuss the evaluation results of the hybrid system we have discussed.
We considered the dataset described in [10] for the experimental evaluation of image spam identification as discussed in Section 4.1, and a variety of emails from the Enron dataset [1], as well as one short text and articles, as detailed in Section 4.2.
Even though we are proposing a hybrid system, we decided to evaluate its two main components, i.e. image and textual spam identification, separately. This decision was taken to allow a more thorough evaluation and comparison with existing non-hybrid approaches.

4.1 Evaluation of Image Spam Recognition

A comparison with other methods is carried out by considering dataset described in [10], which contains 2550 personal ham images, 3239 personal spam images, and 9503 spam Archive images. We set the threshold Th of classification 0.5. The sparse tool introduced in [22] is used to perform sparse representation. To evaluate the performance of our method, we use standard metrics, such as accuracy, precision, recall, and false positive rate (FPR). We use four-fold cross validation to quantitatively analyse the performance. Based on all possible classification results in Table 1, the above metrics can be represented as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{a+d}{a+b+c+d} \\ \text{Precision} &= \frac{a}{a+c} \\ \text{Recall} &= \frac{a}{a+b} \\ \text{FPR} &= \frac{b}{b+d} \end{aligned}$$

To study the roles of local and global features, under the same classification framework, we compare the local features, global features, and their combination. As we can see from Table 2, comparing with single local features or single global features, their combination achieves better classification performance.

Table 2 Comparison of local features, global features, and their combination

	Accuracy	Precision	Recall
Local features	0.953	0.962	0.959
Global features	0.903	0.912	0.909
Local and global features	0.990	0.991	0.989

Table 3 Comparison of LDA and SDA performance

	Accuracy	Precision	Recall
LDA	0.953	0.962	0.959
SDA	0.990	0.991	0.989

Table 4 Comparison of Different Classification Methods

	Accuracy	Precision	Recall
Logistic regression	0.943	0.922	0.929
Support vector machine	0.979	0.982	0.980
Our method	0.990	0.991	0.989

Also, compared with single global features, the performance of single local ones is better, as the former represent the characteristics of a whole image, whereas the latter capture the corresponding details. Although the global features of some spam images are similar to the normal email images, their local features are different. Therefore, it is easier for local features to identify spam.

To validate the contribution of subclass discriminant analysis in feature selection, we have combined local features and global features, and compared linear discriminant analysis (LDA) with subclass discriminant analysis (SDA) to select features in the same classification framework. The final classification performance using LDA and SDA is shown in Table 3. As it can be seen, features selected by SDA can better distinguish spam from normal emails compared to LDA. This is due to the fact that SDA carries out discriminant analysis based on subclasses after clustering, which makes it easier to follow the assumption of Gaussian distribution with a single mode, therefore the selected features by SDA have higher discrimination.

Subsequently, we compare our method with logistic regression and support vector machine, as shown in Table 4.

It can be observed that, as a state-of-the-art classifier, support vector machine is better than logistic regression in terms of accuracy, precision, and recall. However, our method achieves better performance than support vector machine.

Finally, we compare our method with other well-established image spam detection methods, as shown as Table 5. Here, Win et al. [23] employed the histogram and Hough transform to detect spam image. Basheer et al. [24] compared decision tree, bias network, and random forest with texture fea-

Table 5 Quantitative Comparison of our Method with other Methods

	Accuracy	Precision	Recall	False positive
Win et al. [23]	0.954	0.887	0.914	N/A
Basheer et al. [24]	0.985	0.986	0.986	N/A
Zhong et al. [25]	0.920	N/A	N/A	0.010
Our method	0.990	0.991	0.989	0.006

Table 6 A selection of keywords likely to be associated with malicious messages

Clearance
Meet singles
Score with babes
Additional Income
Be your own boss
Double your Salary
Extra income
Income from home
Online degree
Acceptance
Freedom
Hidden
Link
Miracle
Passwords
Satisfaction
Teen
Wife

ture. Zhong et al. [25] employed the texture features based on Wavelet, and applied Active Learning Clustering and feedback-driven semi-supervised support vector machine classification to detect image spams. Table 5 lists the best classification performance of [23] and [24]. It can be observed that, our method achieves better spam detection performance compared to other methods.

4.2 Evaluation of Textual Based Spam Identification

The network G defined in Section 3.2.1, was defined by the semantic structures defined by WordNet. In particular, we only considered the synonymity between words when building the network $H_1 \subset G$. On the other hand, H_2 was manually created as follows

- A team of experts identified a set of keywords likely to be associated with malicious messages, see Table 6 for a selection.
  The edges E_{H_2} were defined based on mutual similarities (again, assessed manually) and keywords that are misspelled, such as "V!AGRA" rather than "VIAGRA". Note that there are typically edges between nodes in V_{H_1} and V_{H_2}. These refer to two words (or keywords) which are synonymous, where on the them is deemed "harmless" , and the other as "malicious",

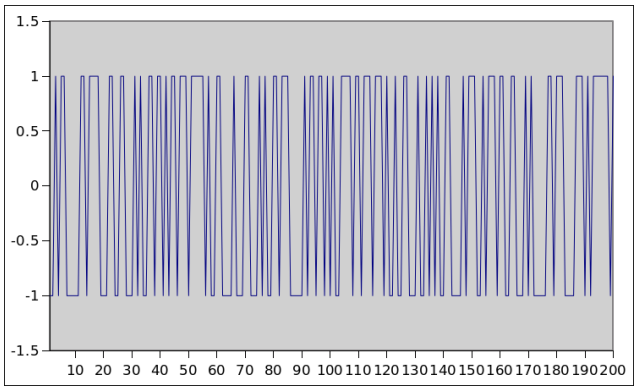

Fig. 6 The diagram depicting the output of the 200 texts discussed in Section 4.2

such as "gender" and "sex". The latter, is more likely to be associated with a harmless message, whereas the former is more likely to be harmless.

We considered the following approach to validate our method

- Approximately 500 emails randomly extracted from the Enron spam dataset, to simulate "static" textual sources. This produced a precision of 96% and a recall of 98%.
- Two articles [27], [28], randomly selected which discuss spam, were analysed. Both texts were identified as non-spam.
- Approximately 200 short texts randomly containing both malicious and genuine texts were also analysed. Each text was added and assessed at sequential time steps, to simulate the dynamic nature of information extraction. In the interpretation of the maliciousness of these texts, we assumed that the trend of the nature of the texts would provide the necessary information, as follows
  - If the analysis of a text shows it has malicious nature, then assume an output equal to -1 is associated with such text.
  - Similarly, if it is non-spam, then the output is 1.
  - If the average of all the outputs is positive, then the texts generate a non-spam document. If it is negative, the created document is spam.

  The above is depicted in Figure 6 . The mean is 0.02, which is very close to 0 and so might be deemed as inconclusive. The manual validation suggested that the generated document is unlikely to be spam, even though only marginally. This is consistent with the output of our method.

5 Conclusion

In this article, a hybrid method to determine whether an unstructured dataset is of malicious nature is discussed. The main motivation is to provide a more flexible and accurate spam detection based on unstructured datasets. As shown

above, experimental results show accuracy and efficiency of our approach, also demonstrating its potential.

In particular, we aim to further expand our approach to fully incorporate a variety of unstructured datasets, not only based on image and textual sources. Furthermore, we are also planning to widen our line of inquiry to fully address non-genuineness. As mentioned, earlier, there are cases where malicious messages have been hidden within seemingly legitimate files (or the other way round), with the intention to be only visible to specific individuals. However, this is by no means a simple task as a deep semantic understanding is required.

Acknowledgements The paper is supported by National Science Foundation of China (61171132), Natural Science Foundation of Jiangsu Province (BK2015022392),Talent Project of Jiangsu Province of China (2014WLW029), Technology Platform Projects of Nantong (CP2013001).

References

1. Nitin, J. and Bing L., Review Spam Detection, Proceedings of the 16th International Conference on World Wide Web, 2007
2. Wertheimer, M. The Mathematics Community and the NSA, Notices of the AMS Volume 62, Number 2, 2015
3. Zhang, C. , et al. , A multimodal data mining framework for revealing common sources of spam images. Journal of Multimedia, 2009. 4(5): p. 313-320.
4. Zhang, C. Image spam clustering: an unsupervised approach. in Proceedings of the First ACM workshop on Multimedia in forensics. 2009: ACM.
5. Mehta, B. , et al. Detecting image spam using visual features and near duplicate detection. in Proceedings of the 17th international conference on World Wide Web. 2008: ACM.
6. Fumera, G. , I. Pillai, and F. Roli, spam filtering based on the analysis of text information embedded into images. The Journal of Machine Learning Research, 2006. 7: p. 2699-2720.
7. Issac, B. and V. Raman. spam detection proposal in regular and text-based image emails. in TENCON 2006. 2006 IEEE Region 10 Conference. 2006: IEEE.
8. Youn, S. and D. McLeod. Improved spam filtering by extraction of information from text embedded image email. in Proceedings of the 2009 ACM symposium on Applied Computing. 2009: ACM.
9. Byun, B. , et al. A Discriminative Classifier Learning Approach to Image Modeling and spam Image Identification. in CEAS. 2007: Citeseer.
10. Dredze, M. , R. Gevaryahu, and A. Elias-Bachrach. Learning Fast Classifiers for Image spam. in CEAS. 2007.
11. Nhung, N. P. and T. M. Phuong. An efficient method for filtering image-based spam. in Research, Innovation and Vision for the Future, 2007 IEEE International Conference on. 2007: IEEE.
12. Trovati M. and Bessis N. An Influence Assessment Method based on Co-Occurrence for Topologically Reduced Big Data Sets, Soft Computing, DOI 10.1007/s00500-015-1621-9, 2015
13. Ahonen, T. , A. Hadid, and M. Pietikainen, Face description with local binary patterns: Application to face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2006. 28(12): p. 2037-2041.
14. Wright, J. , et al. , Robust Face Recognition via Sparse Representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009. 31(2): p. 210-227.
15. Scholkopft, B. and K. -R. Mullert, Fisher discriminant analysis with kernels. Neural networks for signal processing IX, 1999.

16. Zhu, M. and A. M. Martinez, Subclass discriminant analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2006. 28(8): p. 1274-1286.
17. Frey, B. J. and D. Dueck, Clustering by passing messages between data points. science, 2007. 315(5814): p. 972-976.
18. Aharon, M. , M. Elad, and A. Bruckstein, K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. Signal Processing, IEEE Transactions on, 2006. 54(11): p. 4311-4322.
19. Honglak Lee, et al. Efficient sparse coding algorithms. in NIPS. 2006.
20. Kreutz-Delgado, K. , et al. , Dictionary learning algorithms for sparse representation. Neural Comput. , 2003. 15(2): p. 349-396.
21. Menard, S. , Applied logistic regression analysis. Vol. 106. 2002: Sage.
22. Mairal, J. , et al. Online dictionary learning for sparse coding. in Proceedings of the 26th Annual International Conference on Machine Learning. 2009: ACM.
23. Win, Z. M. and N. Aye, Identification of Image spam by Using Histogram and Hough Transform. International Journal, 2013.
24. Al-Duwairi, B. , I. Khater, and O. Al-Jarrah, Detecting Image spam Using Image Texture Features. International Journal for Information Security Research (IJISR), 2012. 2(3/4): p. 344-353.
25. Zhong, J. , Y. Zhou, and W. Deng. Filtering image-based Spam Using Multifractal analysis and active learning feedback-driven semi-supervised support vector machine. in Conference Anthology, IEEE. 2013: IEEE.
26. Kuropka, D. Modelle zur Reprasentation naturlichsprachlicher Dokumente. Logos Verlag, Berlin, 2003
27. Davis, S. and Craney, G., How Do I Stop Spam?, http://www.spamhelp.org/articles/HowDoIStopSpam.pdf
28. Kellett, S. Legislative Definition of Spam for New Zealand, http://www.victoria.ac.nz/law/research/publications/vuwlr/prev-issues/pdf/vol-36-2005/issue-3/kellet.pdf, 2005