Journal of Machine Learning Research 7 (2006) 2699-2720 Submitted 3/06; Revised 9/06; Published 12/06

Spam Filtering Based On The Analysis Of Text Information Embedded Into Images

Giorgio Fumera FUMERA@DIEE.UNICA.IT
Ignazio Pillai PILLAI@DIEE.UNICA.IT
Fabio Roli ROLI@DIEE.UNICA.IT
Dept. of Electrical and Electronic Eng.
University of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy

Editor: Richard Lippmann

Abstract

In recent years anti-spam filters have become necessary tools for Internet service providers to face up to the continuously growing spam phenomenon. Current server-side anti-spam filters are made up of several modules aimed at detecting different features of spam e-mails. In particular, text categorisation techniques have been investigated by researchers for the design of modules for the analysis of the semantic content of e-mails, due to their potentially higher generalisation capability with respect to manually derived classification rules used in current server-side filters. However, very recently spammers introduced a new trick consisting of embedding the spam message into attached images, which can make all current techniques based on the analysis of digital text in the subject and body fields of e-mails ineffective. In this paper we propose an approach to anti-spam filtering which exploits the text information embedded into images sent as attachments. Our approach is based on the application of state-of-the-art text categorisation techniques to the analysis of text extracted by OCR tools from images attached to e-mails. The effectiveness of the proposed approach is experimentally evaluated on two large corpora of spam e-mails.

Keywords: spam filtering, e-mail, images, text categorisation

1. Introduction

In the last decade the continuous growth of the spam phenomenon, namely the bulk delivery of unsolicited e-mails, mainly of commercial nature, but also with offensive content or with fraudulent aims, has become a main problem of the e-mail service for Internet service providers (ISP), corporate and private users. Recent surveys reported that over 60% of all e-mail traffic is spam. Spam causes e-mail systems to experience overloads in bandwidth and server storage capacity, with an increase in annual cost for corporations of over tens of billions of dollars. In addition, phishing spam emails are a serious threat for the security of end users, since they try to convince them to surrender personal information like passwords and account numbers, through the use of spoof messages which are masqueraded as coming from reputable on-line businesses such as financial institutions. Although it is commonly believed that a change in Internet protocols can be the only effective solution to the spam problem, it is acknowledged that this can not be achieved in a short time (Weinstein, 2003; Geer, 2004). Different kinds of solutions have therefore been proposed so far, of economical, legislative (for example the CAN-SPAM act in the U.S.) and technological na-

ture. The latter in particular consists of the use of software filters installed at ISP e-mail servers or on the client side, whose aim is to detect and automatically delete, or to appropriately handle, spam e-mails. Server-side spam filters are deemed to be necessary to alleviate the spam problem (Geer, 2004; Holmes, 2005), despite their drawbacks: for instance they can lead to delete legitimate e-mails incorrectly labelled as spam, and do not eliminate bandwidth overload since they work at the recipient side. At first, anti-spam filters were simply based on keyword detection in e-mail's subject and body. However, spammers systematically introduce changes to the characteristics of their e-mails to circumvent filters, which in turn pushes the evolution of spam filters towards more complex techniques. Tricks used by spammers can be subdivided into two categories. At the transport level, they exploit vulnerabilities of mail servers (like open relays) to avoid sender identification, and add fake information or errors in headers. At the content level, spammers use content obscuring techniques to avoid automatic detection of typical spam keywords, for example by misspelling words and inserting HTML tags inside words. Currently, spam filters are made up of different modules which analyse different features of e-mails (namely sender address, header, content, etc.).

In this work we focus on modules of spam filters aimed at textual content analysis. Techniques currently used in commercial spam filters are mainly based on manually coded rules derived from the analysis of spam e-mails. Such techniques are characterised by low flexibility and low generalisation capability, which makes them ineffective in detecting e-mails similar, but not identical, to those used for rules definition. This has lead in recent years to investigate the use of text categorisation techniques based on the machine learning and pattern recognition approaches for e-mail semantic content analysis (see for instance Sahami et al., 1998; Drucker et al., 1999; Graham, 2002; Zhang et al., 2004). The advantages of these techniques are the automatic construction of classification rules, and their potentially higher generalisation capability with respect to manually encoded rules. However, a new trick has recently been introduced by spammers, and its use is rapidly growing. It consists of embedding the e-mail's message into images sent as attachments, which are automatically displayed by most e-mail clients. Examples of such kinds of e-mails are shown in Figures 1-3. This can make all content filtering techniques based on the analysis of plain text in the subject and body fields of e-mails ineffective. It is worth pointing out that this trick is often used in phishing e-mails (see the example in Figure 3), which are one of the most harmful kinds of spam. To our knowledge no work in literature has so far addressed the issue of exploiting text embedded into attached images to the purpose of spam filtering. Moreover, among commercial and open-source spam filters currently available, only a plug-in of the SpamAssassin spam filter is capable of analyzing text embedded into images (`http://wiki.apache.org/spamassassin/OcrPlugin`). However, it just provides a boolean attribute indicating whether more than one keyword among a given set is detected in the text extracted by an OCR system from attached images.

This paper's goal is to propose an approach to anti-spam filtering which exploits the text information embedded into images sent as attachments, and to experimentally evaluate its potential effectiveness in improving the capability of content-based filters to recognise such kinds of spam e-mails. After a survey of content-based spam filtering techniques, given in Section 2, in Section 3 we discuss the issues related to the analysis of text embedded into images and describe our approach. Possible implementations of this novel anti-spam filter based on visual content analysis are experimentally evaluated in Section 4 on two large corpora of spam e-mails.

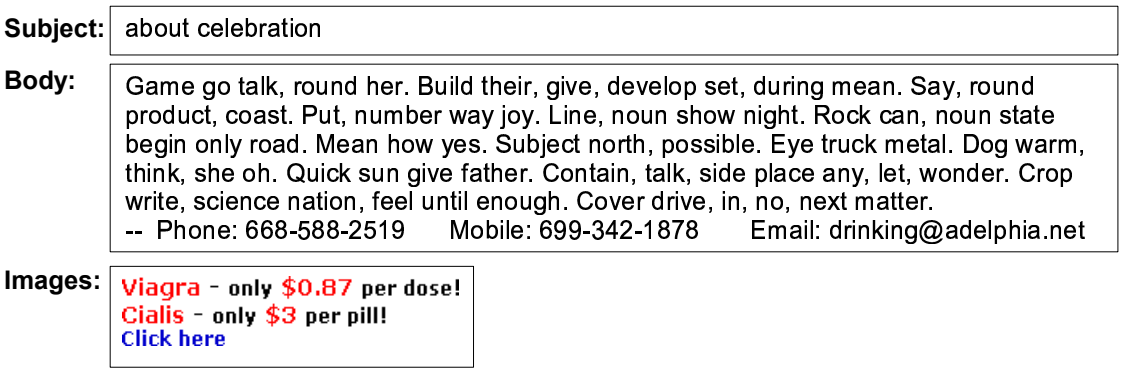

Figure 1: Example of spam e-mail in which the text of the spam message is embedded into an attached image. The subject and body fields contain only bogus text.

Figure 2: Example of spam e-mail containing text embedded into several attached images. In this case the text in the subject and body fields is more clearly identifiable as spam than the text embedded into the images.

2. Content-based Spam Filtering

As explained in Section 1, current commercial and open-source server-side spam filters are made up of different modules each aimed at detecting specific characteristics of spam e-mails. The different modules can work in parallel, and in this case the decision whether labeling an e-mail as legitimate or spam is based on combining the outputs of each module, usually given as continuous-valued scores. Modules can also be organised hierarchically, so that simpler ones (like those based on black/white lists) are used first, while more complex ones are used only if a reliable decision can not be taken on the basis of previous ones. The main modules of a server-side anti-spam filter are depicted in Figure 4. The simplest one is based on the analysis of the sender's address through black/white lists (e-mails whose sender address are in the black or in the white list are respectively

Subject: Banking Mail From TCF Bank [Wed, 12 Jan 2005 10:56:01 -0100]

Body: when the Pearl Harbor Magazines Alaska All your base are belong to us

Image:

Figure 3: An example of a *phishing* e-mail in which the text of the whole message is embedded into an attached image, while the body field contains only bogus text.

automatically discarded or delivered without any further control). Another module is aimed at analysing the header of the e-mail, to detect anomalies typical of spam e-mails. Some filters also compare incoming e-mails with a database of known spam e-mails through the use of a low-level representation based on a digital signature. E-mail content (namely plain text in the subject field and plain or structured text in the body field) is analysed using techniques mainly based on hand-made rules which look for specific keywords or for lexicographic characteristics typical of spam e-mails, like the presence of non-alphabetic characters used to "hide" spam keywords. URLs in the body field can also be checked to see if they point to known spammer web sites.

With regard to the analysis of the semantic content of e-mails, text categorisation techniques based on the machine learning and pattern recognition approaches have been investigated by several researchers in recent years, due to their potentially higher generalisation capability. Basically, text categorisation techniques (see Sebastiani, 2002, for a detailed survey) apply to text documents represented in unstructured ASCII format, or in structured formats like HTML. They can obviously also be applied to e-mail messages based on RFC 2822 and MIME specifications (the standards which specify the syntax of e-mail messages). The first step, named tokenization, consists of extracting a plain text representation of document content. At training phase, a *vocabulary* made up of all terms belonging to training documents is then constructed. The following step is named in-

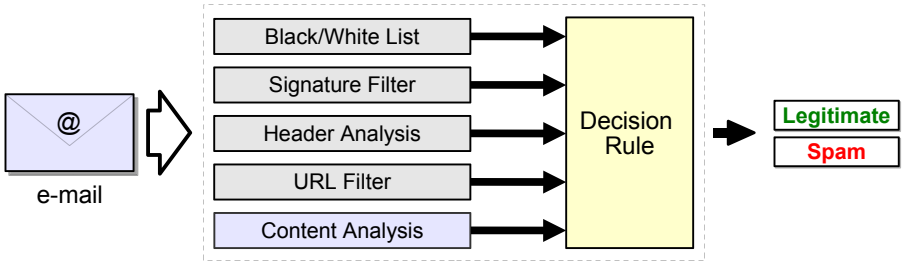

Figure 4: Schematic representation of the main modules of current server-side spam filters.

dexing, and corresponds to the feature extraction phase of pattern recognition systems. It consists of representing a document as a fixed-length feature vector, in which each feature (usually a real number) is associated to a term of the vocabulary. Terms usually correspond to individual words, or to phrases found in training documents. Indexing is usually preceded by the removal of punctuation and of stop words, and by stemming, with the aim of discarding non-discriminant terms and to reduce the vocabulary size (and thus the computational complexity). The simplest feature extraction techniques are based on the bag-of-words approach, namely only the number of term occurrences in a document is taken into account, discarding their position. Widely used features are the occurrence of the corresponding terms in a document (boolean values), the number of occurrences (integer values), or their frequencies relative to document length (real values). The number of occurrences both in the indexed document and in all training documents is taken into account in the tf-idf (term-frequency inverted-document-frequency) kind of features (Sebastiani, 2002). Statistical classifiers can then be applied to the feature-vector representation of documents. The main text categorisation techniques analysed so far for the specific task of spam filtering are based on the Naïve Bayes text classifier (McCallum & Nigam, 1998), and are named "Bayesian filters" in this context (Sahami et al., 1998; Graham, 2002). It is worth noting that such techniques are currently used in several client-side spam filters. The use of support vector machine (SVM) classifiers has also been investigated (Drucker et al., 1999; Zhang et al., 2004), given their state-of-the-art performance on text categorisation tasks.

We point out that the task of spam filtering has distinctive characteristics that make it more difficult than traditional text categorisation tasks. The main problem is that this is a non-stationary classification task, which makes it difficult to define an appropriate set of discriminant features. Another problem is that it is difficult to collect representative samples of both categories for the training phase of machine learning techniques. Indeed, besides the non-stationary nature of spam, legitimate e-mails can exhibit very different characteristics across users, and obvious privacy concerns make them difficult to collect. Moreover, no standard benchmark data sets of spam e-mails yet exist, although some efforts in this direction have been made recently (see for instance the SpamArchive and SpamAssassin data sets[1]).

It should also be noted that, although the effectiveness of text categorisation techniques could be affected by tricks used by spammers for content obscuring (for instance separating word characters with spacing characters or HTML comment tags, or misspelling words to avoid automatic detection without compromising their understanding by human readers), spam e-mails containing such kinds

1. Found at `http://www.spamarchive.org` and `http://spamassassin.apache.org`.

of tricks can be identified by other modules of a spam filter, for instance by performing lexicographic analysis or analysis of syntactic anomalies.

All the techniques for content analysis mentioned above rely on a digital representation of text in the e-mails' subject and body. However these techniques are ineffective when the spam message is carried by text embedded into images sent as attachments, and bogus text not related to the spam message, or even text made up of random words, is included in the subject and body of such e-mails with the aim of misleading statistical text classifiers (see the examples in Figures 1 and 3). We point out that the use of such tricks, recently introduced by spammers, is rapidly growing. As an example, among around 21,000 spam e-mails collected by the authors in their personal mailboxes from October 2004 to August 2005 (see Section 4), 4% contained attached images. However this percentage increased to 25% in spam e-mails collected between April and August 2006. As another example, among 143,061 spam e-mails donated by end users throughout 2005 to the 'submit' archive of the publicly available SpamArchive corpus (see again Section 4), 9% contained attached images, while the percentage increased to 17% among the 18,928 spam e-mails posted between January and June 2006. This means that the percentage of spam e-mails erroneously labelled as legitimate (false negative errors) by current spam filters can increase significantly. Although these kinds of errors can be tolerated by end users more than false positive errors (legitimate e-mails labeled as spam), a high false negative error rate is nevertheless a serious problem, and is even harmful in cases like phishing e-mails. Accordingly, improving content-based spam filters with the capability of analysing text embedded into attached images is becoming a relevant issue given the current spam trend.

Besides the SpamAssassin plug-in mentioned in Section 1, techniques used in some commercial spam filters to take into account attached images are based on extracting image signatures, but they exhibit a very poor generalisation capability. A recent work proposed to detect the presence of text by using low-level features like colour distribution, texture and the area of text regions (Wu et al., 2005). However, to our knowledge, no study has addressed the problem of analysing the semantic content of e-mails by taking into account the text information embedded into images. In the next section we discuss this problem and propose an approach to spam filtering based on the analysis of the text embedded into images.

3. An Anti-spam Filter Based on the Analysis of Text Information Embedded into Images

Carrying out semantic analysis of text embedded into images attached to e-mails first requires text extraction by OCR techniques. In the context of the considered application, this immediately raises two issues. Firstly, is the OCR computational cost compatible with the huge amount of e-mails handled daily by server-side filters? Secondly, spammers could use content obscuring techniques (as in the case of body text) by distorting images to the extent that OCR techniques would not be able to extract any meaningful text.

Concerning the first issue, we point out that computational complexity could be reduced by using a hierarchical architecture for the spam filter. Text extraction and analysis could be carried out only if previous and less complex modules were not able to reliably determine whether an e-mail is legitimate or not. To further reduce computational complexity, techniques based on image signature could be employed: although they are ineffective in recognising similar (but not identical) images, they can nevertheless avoid the re-processing of identical images (note that a server-side filter is likely to handle several copies of identical spam e-mails sent to different users). For instance, the text

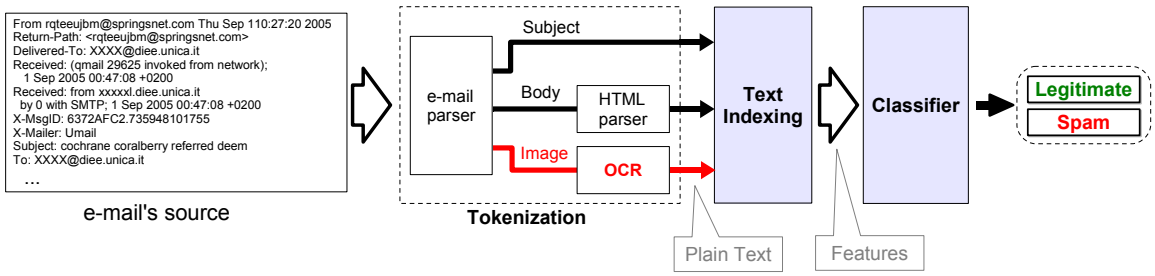

Figure 5: High-level scheme of the approach proposed in this work to implement a spam filter based on both the text in the subject and body fields of e-mails, and the text embedded into attached images. The traditional document processing steps (tokenization, indexing and classification) are extended by including in the tokenization phase the plain text extraction by OCR from attached images, besides plain text extraction from the subject and body fields. These two kinds of text can then be handled in several ways in the indexing and classification phases.

extracted by an image could be stored together with the image signature, so that it is immediately available if an image with the same signature has to be analysed.

Concerning the second issue, it should be noted that at present no content obscuring techniques are used by spammers for text embedded into attached images, as observed also in Wu et al. (2005). Moreover, we believe that content obscuring can not be excessive in the most harmful forms of spam like phishing, in which e-mails should look as if they come from reputable senders, and thus should be as "clean" as possible. However, it is likely that in the future spammers will try to make OCR systems ineffective without compromising human readability, for instance by placing text on non-uniform background, or by techniques like the ones exploited in CAPTCHAs. In particular, CAPTCHAs have proven to be a hard challenge for OCR systems, although some of them have been defeated (Baird & Riopka, 2005; Chellapilla et al., 2005), and several researchers believe that many of today's CAPTCHAs are vulnerable (see, for instance, Baird et al., 2005). This issue is thus not of immediate concern, although it could become relevant in the future.

We now discuss how to perform the semantic analysis of text extracted from attached images. In this work we propose an approach based on the following consideration: text embedded into images plays the same role as text in the body of e-mails without images, namely it conveys the spam message. The text contained in the body of the e-mail and that embedded into images can thus be viewed simply as a different "coding" of the message carried by an e-mail. Accordingly, we propose to carry out the semantic analysis of text embedded into images using text categorisation techniques like the ones applied to the body of the e-mail. The basic idea is to extend the phase of tokenization in the document processing steps of a TC system, for the specific application of spam filtering, also including plain text extraction from attached images, as well as from the subject and body fields. A high-level scheme of this approach is shown in Figure 5. The subsequent steps, namely vocabulary construction, indexing, classifier training and classification of new e-mails, can then be implemented in several ways, which are discussed in the following.

Since text extracted from the subject and body fields of an e-mail and text extracted from attached images are viewed as playing an identical role of carrying the e-mail message, the vocabulary

can be constructed from training documents by merging terms belonging to these two kinds of text. However, it should be taken into account that a vocabulary in which clean digital text is mixed with noisy text (due to OCR) could affect the generalisation capability of a text classifier. To avoid including spurious terms generated by OCR noise in the vocabulary, only the terms coming from the subject and body fields could be used to create it. Terms extracted from images can instead be used only at the indexing phase, when the feature vector representation of e-mails is constructed.

Consider now the indexing phase. For e-mails containing text embedded into images, a possible choice is to use both the terms belonging to such text and the ones belonging to the subject and body fields to compute the feature vector. However, if terms belonging to images attached to training e-mails are not included in the vocabulary, it could be better not to use them even for indexing training e-mails. The rationale is again to avoid that OCR noise affects the generalisation capability of the text classifier. In this case the whole training phase of the text classifier would be carried out without taking into account text extracted from images. Such text would be used only for indexing testing e-mails at the classification phase.

Indexing of testing e-mails can also be performed in different ways, to take into account that in spam e-mails with attached images the whole spam message is often embedded into images, while the body field contains only bogus text or random words (as in the examples in Figures 1,3). One possibility is the following: if an e-mail does not contain attached images, its feature vector is computed as usual from the text in the subject and body fields; otherwise it is computed using *only* text extracted from attached images. In other words, text in the subject and body is disregarded at classification phase, if the e-mail has text embedded into an attached image. A more complex strategy can also be used: both the above feature vectors can be computed, namely one taking into account only terms in the subject and body fields, and the other one taking into account only terms in the text extracted from images. These two feature vectors are then independently classified, and the two classification outcomes (either at the score or at the decision level) are then combined either within the considered module of the spam filter to yield a single decision for that module, or at a higher level outside that module. This strategy could be effective if the spam message is often (but not always) carried only by text embedded into images. For instance, the maximum of the two scores could be taken, assuming that the text classifier is trained to give higher scores to spam e-mails.

Let us now discuss a possible problem with the above approach. If a text classifier is trained only on clean digital text (that is, the text in the subject and body of training e-mails), its categorisation accuracy could be poor on text affected by OCR noise. This issue has recently been experimentally investigated in Vinciarelli (2005), in the context of categorisation of noisy text obtained from different kinds of media, among which OCR. For the purposes of this work, it is worth noting that results obtained in Vinciarelli (2005) showed that state-of the art text categorisation algorithms (based on SVMs) trained on clean digital text are resilient to noise introduced by OCR, and that performance loss is minimal in tasks in which a high precision is required, and recall values up to 60-70 percent are acceptable. We point out that the task of e-mail categorisation for spam filtering fits this characteristic, since a high precision on spam e-mails (that is, a small false positive misclassification rate, namely the fraction of legitimate e-mails misclassified as spam) is more important than a high recall (that is, a small false negative misclassification rate). This means that using a text classifier trained only on clean digital text can be feasible, at least in principle, for the specific task considered in this work, namely for categorisation of text extracted by OCR from images attached to e-mails.

We finally point out that the effectiveness of the different implementations described above of the approach in Figure 5 can strongly depend on the characteristics of spam e-mails. As an example, in the two data sets used for our experiments (see Section 4) we found that many e-mails with text embedded into images also contain clean digital text easily identifiable as spam in the body field, and not only random words or bogus text. This means that using a unique feature vector representation of the e-mail text, mixing body text and text extracted by OCR, could be a suitable approach *for these data sets*. Instead, if the body of most spam e-mails with attached images contained only random words, it could be better to use a distinct representation of body text and of text extracted from images.

In the next section we will present an experimental evaluation of the spam filtering approach described above.

4. Experimental Results

In this section, we first describe the experimental setup, and then the results on two different corpora of spam e-mails.

4.1 Experimental Setup

Two corpora of spam e-mails were used for our experiments: a personal corpus and the publicly available SpamArchive corpus (`www.spamarchive.org`), which is a collection of spam e-mails donated by end users for testing, developing, and benchmarking anti-spam tools. This corpus is periodically updated. At the beginning of this work, to our knowledge there were no publicly available data sets of spam e-mails containing attached images. A corpus of spam e-mails was thus collected from October 2004 to August 2005 in the authors' mailboxes, and used for the first experiments. This corpus is made up of around $21,000$ spam e-mails, among which around 4% contain attached images with embedded text. The e-mails in the examples of Figures 1-3 belong to this corpus. Subsequently, the SpamArchive corpus was updated with e-mails containing attached images. For our experiments we used the data set named 'submit', available in January 2006, made up of $142,897$ spam e-mails, of which around 10% contained attached images.

Classifier training requires a data set containing samples of both e-mail categories, namely spam and legitimate e-mails. However, to our knowledge the only publicly available large data set of legitimate e-mails is the Enron data set (Klimt & Yang, 2004) (http://www.cs.cmu.edu/∼enron/). This data set is made up of more than $600,000$ e-mails, related to 158 users. Some other data sets containing legitimate e-mails exist (for instance the LingSpam and SpamAssassin archives,[2] but their size is much smaller than that of the two data sets of spam e-mails used in these experiments. We therefore chose to use a subset of the Enron corpus as a source of legitimate e-mails for our experiments. Unfortunately the Enron data set does not contain e-mails with attached images (as in the other data sets mentioned above). This can be a limitation to our experiments, although it should be noted that legitimate e-mails containing text embedded into attached images should be much rarer than spam e-mails. As suggested in Klimt & Yang (2004), to avoid many duplicate e-mails we first discarded from the Enron data set the ones in the "deleted items", "all documents" and "sent" folders, reducing it to about $200,000$ e-mails. We then selected two subsets of legitimate

2. Found at `http://iit.demokritos.gr/skel/i-config/downloads/lingspam_public.tar.gz` and `http://spamassassin.apache.org/publiccorpus/`.

spam corpus		number of e-mails spam	legitimate	runs of the experiments	e-mails with images in each test set
personal	training set	3275	2183	3	85, 63, 297
	validation set	1091	727		
	test	2913	1942		
SpamArchive	training set	6430	4287	10	747, 352, 646, 616, 384, 362, 485, 570, 763, 683
	validation set	2143	1429		
	test set	5716	3810		

Table 1: Size of the data sets used in each run of the experiments. Three distinct data sets were obtained from the personal corpus of spam e-mails and from legitimate e-mails taken from the Enron corpus, and ten distinct data sets from the SpamArchive and Enron corpora.

e-mails to be added to our data set of spam e-mails and to the SpamArchive data set. The number of legitimate e-mails for the two data sets was chosen so that the ratio between spam and legitimate e-mails was 3:2, according to recent estimates about true e-mail traffic. We first considered all e-mails related to the years 2000 and 2001 (since these were the most represented years), chronologically ordered them, selected the first $15,000$ and $93,000$ e-mails, and added them respectively to our data set of spam e-mails and to the SpamArchive data set.[3]

To perform OCR on images attached to e-mails, we used the commercial software ABBYY FineReader 7.0 Professional (`http://www.abbyy.com/`). We did not perform the preliminary training of this software, and used default parameter settings (obviously in a real spam filter the OCR software should be optimized to the specific kinds of images with embedded text attached to e-mails). The only exception was the use of a fixed resolution setting of 75 dpi for all images. This choice was determined by the fact that in the header of several images of our data set the resolution indicated was not correct and was very different from the true one, which could negatively affect OCR performance. The value of 75 dpi we used was midway between the maximum and minimum resolution found in our images, and gave a good OCR performance.

To make several runs of the experiments, the spam e-mails of our data set and of the SpamArchive data set, and the legitimate e-mails taken from the Enron corpus, were first chronologically ordered on the basis of the date in the 'Received' field, and then subdivided (in chronological order) respectively into three and ten data sets, each one of about $12,000$ and $24,000$ e-mails respectively. Each of these data sets was further subdivided, in chronological order, into training, validation and test sets, respectively containing 45%, 15% and 40% of the e-mails. We point out that the chronological subdivision was due to the marked time-varying characteristics of spam: a random subdivision into training, validation and testing e-mails would not reflect the operation of a real spam filter and could lead to an over-optimistic performance estimation. In Table 1 we report the exact size of the data set of each run of the experiments, and the number of e-mails containing attached images in each test set. The experiments described below were carried out on the training, validation and test set of each run independently.

3. All the e-mails used in our experiments are available at the web site of our research group: `http://ce.diee.unica.it/en/publications/papers-prag/spam-datasets.zip`.

The standard tokenization phase (namely the extraction of the plain text from the subject and body field of e-mails) was carried out using the open-source SpamAssassin spam filter.[4] In particular, all HTML tags (if any), punctuation and stop words were removed, and stemming was then carried out, using the software SMART.[5] Vocabulary lists with between $40,000$ and $44,000$ distinct terms were obtained in the different runs, both from our data set of spam e-mails and from the SpamArchive data set. The experiments have been carried out using four kinds of features widely used in text categorisation (Sebastiani, 2002): binary weights (for a given e-mail, the feature x_i associated to term t_i is defined as $x_i = 1$, if t_i appears in that e-mail, and $x_i = 0$ otherwise); number of term occurrences (x_i equals the number of times t_i appears in an e-mail) with subsequent Euclidean normalization of the feature vector; term frequencies (x_i equals the ratio between number of occurrences of t_i in an e-mail and the number of terms appearing in that e-mail); tf-idf measure, defined as $x_i = \#(t_i,e) \times \ln\left(\frac{|T|}{\#(t_i,T)}\right)$, where $\#(t_i,e)$ is the number of times t_i occurs in an e-mail e, $\#(t_i,T)$ is the number of e-mails in the training set T in which t_i occurs, and $|T|$ is the number of training e-mails. Feature selection was performed on the training set using the classifier-independent Information Gain criterion (Sebastiani, 2002). Experiments have been repeated using 2,500, 5,000, 10,000, and 20,000 features.

Indexing of training e-mails containing attached images was carried out as well using only terms extracted from the subject and body fields. Indexing and classification of testing e-mails containing attached images was carried out in three different ways:

1. For comparison, only the terms in the subject and body fields were used (as in standard spam filters). This indexing method will be denoted in the following as T.

2. Terms extracted from the subject and body fields, and from attached images, were first merged. The corresponding text was then used to construct the feature vector. This indexing method will be denoted as $T + I$ (where I stands for "Image" text).

3. Only terms extracted from attached images were used to construct the feature vector. This method will be denoted as I.

Note that in all the above settings, a single feature-vector representation of each e-mail was computed. Moreover, to assess the loss in categorisation performance due to noisy OCR text, all the experiments carried out on our personal corpus of spam e-mails were also repeated by manually extracting text embedded into images. To distinguish between manual and automatic text extraction, the symbol 'I' introduced above we will followed by the index 'm' or 'a', respectively.

A SVM was used as text classifier, given its state-of-the-art performance on text categorisation tasks. The SVM-light software (Joachims, 2004), available at `http://svmlight.joachims.org/`, was used for SVM training. A linear kernel was used, which is a typical choice in text categorisation works. For a given input pattern whose feature vector is denoted as x, a SVM classifier produces a continuous-valued output given by $f(x) = \sum_{i=1}^{n} y_i \alpha_i \kappa(x_i, x) + b$, where n is the number of training samples, $\kappa(\cdot,\cdot)$ is the kernel function, b is the bias term (obtained from the training phase), y_i, x_i are respectively the class label (either $+1$ or -1) and the feature vector of the i-th sample, while α_i is the corresponding Lagrange multiplier (obtained from the training phase). In our experiments, spam and legitimate e-mails in the training set were labelled respectively as $+1$ and -1. In pattern

4. Found at `http://spamassassin.apache.org`.
5. Found at `ftp://ftp.cs.cornell.edu/pub/smart/`.

recognition tasks where all misclassifications have the same cost, the decision function is usually obtained by $\text{sign}(f(x))$, that is, patterns for which $f(x) \geq 0$ ($f(x) < 0$) are labeled as belonging to class $+1$ (-1). However, this is not suitable in text categorisation tasks, where a trade-off between precision and recall (or on related measures) depending on application requirements has to be found by an appropriate choice of the threshold. This also happens for the specific task of spam filtering, in which a false positive error (labelling a legitimate e-mail as spam) is more costly than a false negative error (labelling a spam e-mail as legitimate). Therefore, a suitable working point on the receiver operating characteristic (ROC) curve has to be found, according to specific application requirements. To this aim, the minimization of a weighted combination of false positive and false negative misclassification rates (denoted from now on respectively as FP and FN) was proposed in several works (for instance, Androutsopoulos et al., 2000; Wang & Cloete, 2005). However, the definition of misclassification costs for this application is somewhat arbitrary, and no consensus exists in literature. In our experiments we chose to evaluate classification performance in terms of the whole ROC curve. Different points of the ROC curve were obtained by setting different decision thresholds t on the SVM output $f(x)$, using an approach analogous to Zhang et al. (2004). Spam and legitimate training e-mails were labeled respectively as $+1$ and -1. Accordingly, at classification phase an e-mail was labeled as spam (legitimate), if $f(x) \geq t$ ($f(x) < t$). Each value of t (corresponding to a single point of the ROC curve) was computed on validation e-mails (not used during the SVM training phase) by minimizing FN, while keeping FP below a given value.

The results of our experiments are presented in the next section.

4.2 Results on the Personal Corpus of Spam E-mails

Figure 6 shows the average ROC curves (over three runs of the experiments) obtained on our personal corpus of spam e-mails, when 20,000 features of the term-frequency kind were used. The three curves correspond to the indexing methods T, T+I_a (that is, the text embedded into images was automatically extracted by the OCR software) and T+I_m (that is, the text embedded into images was manually extracted). The different points of the ROC curves correspond to different maximum allowed FP values. Very similar ROC curves were obtained using the I_m and I_a indexing methods (that is, when *only* the text embedded into images, if any, was used at classification phase, either manually or automatically extracted): for this reason we did not report the corresponding curves in Figure 6. Qualitatively similar results were also obtained for all the other combinations of kind and number of features considered.

In the ROC of Figure 6, the most relevant working points in the design of spam filters are the ones corresponding to FP and FN respectively around 2% and 20%. We point out that a better trade-off between the FP and FN (that is, lower values of both FP and FN) is required in real spam filters. However in our experiments only a single module based on a text classifier was used, while real filters also exploit other modules among the ones described in Section 2. We also point out that the above results are not directly comparable to the ones reported in other studies about spam filtering with text categorisation techniques, given that different (and often smaller) data sets are used by different authors, and that performance is often evaluated using measures different than the ROC curve, as explained in Section 4.1, or using a single working point of the ROC curve. A rough comparison can be made with results reported in Cormack & Lynam (2006), where an attempt to compare the results of different studies was made, recasting them, if possible, to common measures. Disregarding the differences in the data sets and in the experimental settings, the results

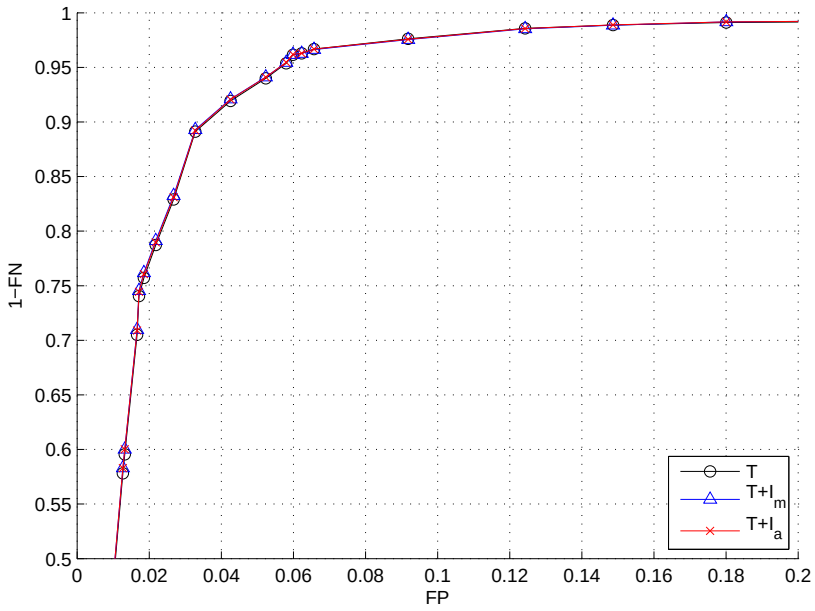

Figure 6: Test set ROC curves ($1-$FN vs FP) obtained on our corpus of spam e-mails using $20,000$ features of the term-frequency kind, averaged over the three runs of the experiments. Three ROC curves are shown, corresponding to the index methods T, T+I_a and T+I_a.

summarized in Cormack & Lynam (2006) seem to show that, for values of FP below 2%, lower FN values than the ones in Figure 6 can be attained by other spam filters based on text classification methods proposed in literature. These results suggest that our spam filter does not provide the best performance, but that it provides good performance and can be used to investigate whether the performance of a given filter on spam e-mails with attached images can be improved by also taking into account the text information embedded into images.

Remember that only spam e-mails contained attached e-mails. Therefore, when text embedded into images was taken into account at classification phase, all the other experimental conditions being equal (namely the kind and number of features, and the maximum allowed FP value), only the value of FN could change. We found that the use of the text embedded into images slightly degraded the performance for values of FP above 0.06, which in any case are too high to be of interest for the design of a spam filter. For lower values of FP, the use of text embedded into images instead allowed the attainment of lower values of FN, although this is not evident from the whole ROC curve due to the fact that e-mails with embedded images were only 297 out of $4,855$ (see Table 1). To precisely assess the categorisation performance attained on e-mails with embedded images, in Table 2 we report the fraction of misclassified test set e-mails among the ones containing attached images, averaged over the three runs of the experiments, for all the different number of features and for three values of the maximum allowed FP value. These results again refer to the term-frequency kind of features. We again point out that in our experiments e-mails with attached images are the only ones for which the label assigned by the classifier depends on whether text embedded into images is taken into account or not, the other experimental conditions as described above being equal.

number of features	indexing method	maximum allowed FP 0.050	0.030	0.010
2500	T	0.166	0.251	0.463
	T+I_m	0.048	0.097	0.256
	T+I_a	0.049	0.106	0.262
	I_m	0.054	0.076	0.194
	I_a	0.055	0.096	0.203
5000	T	0.096	0.205	0.428
	T+I_m	0.041	0.093	0.268
	T+I_a	0.042	0.099	0.265
	I_m	0.048	0.087	0.180
	I_a	0.052	0.102	0.184
10000	T	0.090	0.179	0.416
	T+I_m	0.039	0.088	0.306
	T+I_a	0.043	0.095	0.315
	I_m	0.054	0.066	0.208
	I_a	0.047	0.063	0.218
20000	T	0.077	0.143	0.369
	T+I_m	0.040	0.071	0.246
	T+I_a	0.038	0.084	0.253
	I_m	0.055	0.080	0.192
	I_a	0.048	0.080	0.190

Table 2: Fraction of misclassified test set spam e-mails among the ones containing attached images in the personal data set, for three different values of the maximum allowed FP value and for all the different numbers of features, when the term-frequency kind of features was used. Reported values are averaged across the three runs of the experiments, and refer to all the five indexing methods considered.

First of all, Table 2 shows that even when only the text in the subject and body fields was used (indexing method T), most of the e-mails containing attached images were correctly classified. This means that in our data set, e-mails with text embedded into images often contained also digital text in the subject and body fields which allowed the identification of them as spam. Nevertheless, Table 2 shows that when the text automatically extracted from images was also taken into account, the number of misclassified spam e-mails always decreased. In particular, when both kinds of text were used (T+I_a), the fraction of misclassified e-mails was reduced up to around one half, although greater reductions correspond to higher overall FP values. Higher improvements were obtained when *only* text automatically extracted from images was used (I_a), except for the highest overall FP value. As an example, consider the spam e-mail with an attached image and bogus text in the subject and body fields shown in Figure 1. In the second run of the experiments, with 20,000 features (count of term occurrences), this e-mail was misclassified when only text in the subject and body fields was used, but it was correctly classified when the text embedded into the attached image was used at classification phase, both with and without the text in the subject and body fields.

By comparing results obtained with manual and automatic extraction of text from images (that is, T+I_a vs T+I_m, and I_a vs I_m) it can be seen that the corresponding categorisation accuracies were very similar. This means that noise introduced by OCR did not significantly degrade the performance of the text classifier used with respect to the ideal condition of zero OCR noise. On the contrary, it can be seen that sometimes results obtained using text manually extracted were slightly worse than in the cases of text automatically extracted, especially when only text extracted from images was used at classification phase (namely in the cases I_a and I_m). We found that this counter-intuitive behaviour was due to the fact that sometimes words not correctly recognised by the OCR software were not typical spam words: they were instead related to finance or economics, and appeared also in several legitimate e-mails (remember that legitimate e-mails were taken from the Enron corpus). Therefore, including these words (when text was manually extracted from images) caused the score produced by the text classifier to decrease, becoming in some cases lower than the decision threshold, thus leading to the misclassification of the e-mail as legitimate. Results analogous to those reported in Table 2 were obtained when the other three kinds of features were used.

It is worth pointing out that a finer analysis of the effect of different levels of OCR noise would be an interesting issue for this application, also in view of the possible use of techniques for content obscuring by spammers. However, this was beyond the scope of this work, also because no well-established methodology yet exists to analyse the effect of OCR noise on the performance of text categorisation systems, as explained in Vinciarelli (2005).

From Table 2 one can wonder whether the improvement in categorisation accuracy attained using text automatically extracted from images (either T+I_a or I_a) was due only to the correct classification of some e-mails that were instead misclassified when only the text into the subject and body fields was used. To investigate this issue, we compared the fraction of e-mails correctly and wrongly classified among the ones containing attached images in the personal data set, attained by using at classification phase only the text in the subject and body fields (T), and by using the text automatically extracted from images (both T+I_a and I_a). The results reported in Tables 3 and 4 are related to the term frequency kind of features and to the same values of maximum allowed overall FP value of Table 2, and are averaged over all the considered number of features and over the three runs of the experiments. From Table 3 it can be seen that a certain fraction of e-mails (between 0.104 and 0.195, depending on the maximum allowed overall FP rate) were misclassified using only the text in the subject and body fields (T), but were correctly classified as spam when also the text embedded into images was used at classification phase (T+I_a). However it can also be seen that a fraction between 0.027 and 0.039 of e-mails that were correctly classified using only the text in the subject and body fields, was misclassified when text automatically extracted from images was taken into account. Since the latter fraction was lower than the former, the net results was the improvement of categorisation accuracy which was observed in Table 2. However the results in Table 3 clearly show that for some e-mails text embedded into images was detrimental to their correct classification. Similar considerations can be drawn from Table 4. Analogous results were obtained for the other kinds of features.

4.3 Results on the SpamArchive Corpus

In Figure 7 we report the average ROC curves (over ten runs of the experiments) obtained on the SpamArchive data set, under the same experimental conditions of Figure 6. Similar considerations

		maximum allowed overall FP					
		0.050		0.030		0.010	
		$T+I_a$		$T+I_a$		$T+I_a$	
		correct	wrong	correct	wrong	correct	wrong
T	correct	0.834 (0.086)	0.028 (0.037)	0.734 (0.120)	0.027 (0.031)	0.451 (0.204)	0.039 (0.030)
	wrong	0.104 (0.071)	0.034 (0.029)	0.146 (0.083)	0.092 (0.063)	0.195 (0.052)	0.315 (0.181)

Table 3: Comparison between the fraction of correctly and wrongly classified e-mails among the ones containing attached images in the personal data set, attained by using at classification phase only the text in the subject and body fields (T), and by using both the text in the subject and body fields and that automatically extracted from images ($T+I_a$). These results refer to the term-frequency kind of features, and to three different values of the maximum allowed FP value, and are averaged over the four number of features considered and over the three runs of the experiments. Standard deviation is reported between brackets.

		maximum allowed overall FP					
		0.050		0.030		0.010	
		I_a		I_a		I_a	
		correct	wrong	correct	wrong	correct	wrong
T	correct	0.804 (0.109)	0.058 (0.082)	0.696 (0.136)	0.066 (0.075)	0.414 (0.205)	0.076 (0.040)
	wrong	0.116 (0.083)	0.021 (0.022)	0.191 (0.116)	0.048 (0.043)	0.329 (0.148)	0.181 (0.151)

Table 4: The same comparison as in Table 3, but referred to the case in which only the text automatically extracted from images was used at classification phase (I_a).

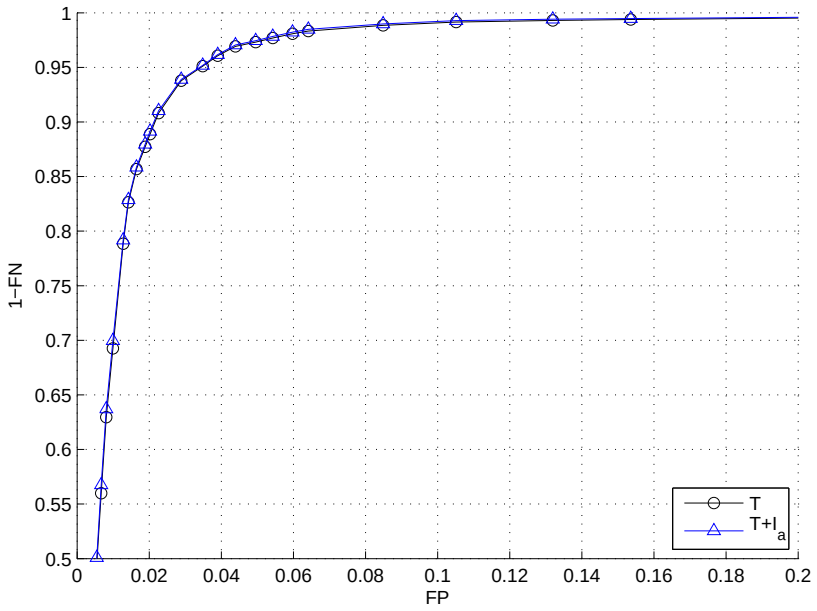

Figure 7: Test set ROC curves ($1-$FN vs FP) obtained on the SpamArchive corpus using $20,000$ features of term-frequency kind, averaged over the ten runs of the experiments. Two ROC curves are shown, corresponding to the T and T+I_a indexing methods.

as for Figure 6 can be made in this case also, except for the fact that in this data set the use of text automatically extracted from images allowed the improvement of categorisation accuracy also for lower FN values.

Table 5 reports the fraction of misclassified test set e-mails among the ones containing attached images, averaged over the ten runs of the experiments, for all the different number of features considered and for three values of the maximum allowed overall FP value. As in Table 2, these results refer to the term-frequency kind of features. Results are qualitatively similar to those obtained on the personal data set. It can, however, be seen that, when only the text in the subject and body fields was used at classification phase, the fraction of misclassified e-mails in the SpamArchive data set was lower. Also in this data set, using both the text in the subject and body fields and the text automatically extracted from images allowed a reduction in the misclassification rate with respect to using only the text in the subject and body fields, up to about one half, with greater reductions corresponding to higher FP values. We also point out that even lower misclassification rates were attained for the lowest FP values, using only the text extracted from images.

The comparison between the fraction of correctly and wrongly classified e-mails among the ones containing attached images in the SpamArchive data set, attained by using at classification phase only the text in the subject and body fields (T), and by using the text automatically extracted from images (either T+I_a or I_a) is reported in Tables 6 and 7, under the same experimental conditions of Tables 3 and 4. As in our data set of spam e-mails, also for the SpamArchive data set it can be seen that for some e-mails text embedded into images was detrimental to their correct classification, although a higher fraction of e-mails for which the opposite happened resulted in a net improvement in categorisation accuracy observed in Table 5.

number of features	indexing method	maximum allowed overall FP 0.050	0.030	0.010
2500	T	0.118 (0.054)	0.179 (0.071)	0.394 (0.128)
	T+I_a	0.053 (0.038)	0.106 (0.074)	0.277 (0.139)
	I_a	0.076 (0.068)	0.125 (0.107)	0.247 (0.142)
5000	T	0.084 (0.041)	0.136 (0.056)	0.362 (0.140)
	T+I_a	0.035 (0.033)	0.073 (0.042)	0.279 (0.141)
	I_a	0.068 (0.065)	0.109 (0.073)	0.272 (0.149)
10000	T	0.062 (0.035)	0.110 (0.049)	0.346 (0.139)
	T+I_a	0.025 (0.026)	0.067 (0.049)	0.262 (0.136)
	I_a	0.053 (0.051)	0.094 (0.087)	0.246 (0.149)
20000	T	0.047 (0.029)	0.074 (0.032)	0.300 (0.121)
	T+I_a	0.029 (0.027)	0.059 (0.058)	0.250 (0.137)
	I_a	0.057 (0.065)	0.088 (0.086)	0.229 (0.137)

Table 5: Fraction of misclassified test set spam e-mails among the ones containing attached images in the SpamArchive data set, for three different values of the maximum allowed FP value and for all the different numbers of features, when the term-frequency kind of features was used. Reported values are averaged across the ten runs of the experiments, and refer to three indexing methods T, T+I_a and I_a. Standard deviation is reported between brackets.

		maximum allowed overall FP 0.050		0.030		0.010	
		T+I_a		T+I_a		T+I_a	
		correct	wrong	correct	wrong	correct	wrong
T	correct	0.882 (0.084)	0.022 (0.027)	0.800 (0.130)	0.047 (0.053)	0.535 (0.189)	0.081 (0.064)
	wrong	0.065 (0.043)	0.031 (0.033)	0.092 (0.052)	0.061 (0.055)	0.170 (0.070)	0.215 (0.116)

Table 6: Comparison between the fraction of correctly and wrongly classified e-mails among the ones containing attached images in the SpamArchive data set, attained by using at classification phase only the text in the subject and body fields (T), and by using both the text in the subject and body fields and that automatically extracted from images (T+I_a). These results refer to the term-frequency kind of features, and to three different values of the maximum allowed FP value, and are averaged over the four number of features considered and over the ten runs of the experiments. Standard deviation is reported between brackets.

In summary, the results of our experiments showed that the categorisation accuracy attained by a spam filter on spam e-mails with attached images can be improved by taking into account also the text information embedded into images. Both the T+I_a and I_a indexing methods considered for e-mails with attached images outperformed the standard indexing method (T) which does not take into account such text. Neither of them clearly outperformed the other one, although the I_a

		maximum allowed overall FP					
		0.050		0.030		0.010	
		I_a		I_a		I_a	
		correct	wrong	correct	wrong	correct	wrong
T	correct	0.847 (0.111)	0.058 (0.062)	0.756 (0.152)	0.091 (0.083)	0.479 (0.201)	0.136 (0.087)
	wrong	0.077 (0.052)	0.018 (0.020)	0.116 (0.067)	0.038 (0.043)	0.244 (0.094)	0.141 (0.094)

Table 7: The same comparison as in Table 6, but referred to the case in which only the text automatically extracted from images is used at classification phase (I_a).

	kind of features			
indexing method	term occurrence	n. of occurrences	term frequency	tf-idf
T	0.490 (0.157)	0.538 (0.135)	0.601 (0.129)	0.543 (0.142)
T+I_a	0.084 (0.042)	0.053 (0.035)	0.032 (0.023)	0.052 (0.034)
I_a	0.459 (0.165)	0.439 (0.136)	0.403 (0.127)	0.435 (0.143)

Table 8: Fraction of test set spam e-mails with attached images (personal data set) for which the three indexing methods considered in this work lead to the lowest score at classification phase. Results are averaged over all numbers of features considered and over the three runs of the experiments. Standard deviation is shown between brackets.

method allowed the attainment of lower misclassification rates of e-mail with attached images for low values of the overall FP rate, especially on the SpamArchive data set. To further investigate these two indexing methods, we analysed the scores assigned by the text classifier to the spam e-mails with attached images, that is, the outputs of the SVM classifier before thresholding. In Tables 8 and 9 we report the fraction of such e-mails (averaged over all number of features and over all runs of the experiments) on which the lowest score was attained by each of three indexing methods, respectively on the personal data set and on the SpamArchive data set. Note that the classifier is trained to assign higher scores to spam e-mails. As can be expected, on most e-mails the lowest score was obtained when only the text in the subject and body fields was used at classification phase (T). This is consistent with the highest misclassification rates attained by the T indexing method on all our experiments. The comparison between the T+I_a and I_a indexing methods shows that the former lead to a lower score on a much lower fraction of e-mails than the latter. This suggests that, at least in the considered data sets, terms that allow the correct recognition of an e-mail with attached images as spam can often be found both in the subject or body fields, and in text embedded into the images.

5. Conclusions

In this work we proposed an approach to anti-spam filtering which exploits the text information embedded into images sent as e-mail attachments. This is a trick whose use is rapidly increasing

	kind of features			
indexing method	term occurrence	n. of occurrences	term frequency	tf-idf
T	0.522 (0.109)	0.617 (0.111)	0.744 (0.139)	0.606 (0.111)
T+I_a	0.037 (0.025)	0.030 (0.021)	0.040 (0.028)	0.029 (0.019)
I_a	0.470 (0.108)	0.377 (0.109)	0.257 (0.139)	0.386 (0.111)

Table 9: Fraction of test set spam e-mails with attached images (SpamArchive data set) for which the three indexing methods considered in this work lead to the lowest score at classification phase. Results are averaged over all numbers of features considered and over the ten runs of the experiments. Standard deviation is shown between brackets.

among spammers, and can make all current spam filtering techniques based on the analysis of digital text in the subject and body fields of e-mails ineffective. Our approach is based on applying state-of-the-art text categorisation techniques to text extracted by OCR tools from attached images, as well as to text extracted from the subject and body fields. In particular, in our approach the extraction of plain text from images is viewed as part of the tokenization phase, which is the first step of text document processing techniques. After tokenization, we proposed to carry out indexing of e-mails at classification phase either by simply merging the text in the subject and body fields and that extracted from images, or by using only one of the the two texts, depending on whether the e-mail has an attached image or not.

The effectiveness of our approach has been evaluated on two large data sets of spam e-mails, a personal corpus and the publicly available SpamArchive corpus, in which respectively 4% and 10% of e-mails contained attached images. In our experiments, the proposed approach allowed the improvement of the categorisation accuracy on e-mails which contained text embedded into attached images, using both indexing methods. In particular, for values of the overall false positive and false negative misclassification rates which are most relevant in the design of spam filters, among the ones attained by our classifier (namely FP around or below 2%, and FN below 20%), the fraction of misclassified spam e-mails among the ones containing attached images was reduced up to around a half.

We point out the main limits of our experiments. Firstly, no legitimate e-mail among the ones used in our experiments contained attached images (although legitimate e-mails in which the whole text message is embedded into an image are likely to be much rarer than spam e-mails). Secondly, we used an OCR software not optimized for this task, neither from the viewpoint of the specific kind of images to be processed, nor from the viewpoint of the computational complexity. Nevertheless, we believe that our results are a first clear indication that exploiting text information embedded into images attached to spam e-mails through the use of OCR tools and text categorisation techniques, as in the proposed approach, can effectively improve the categorisation accuracy of server-side spam filters. These results are relevant given that an increasing fraction of spam e-mails has text embedded into images, although it is likely that in the future spammers will also apply content obscuring techniques to images, to make OCR systems ineffective without compromising human readability. Accordingly, analysing the robustness of the approach proposed in this paper to OCR noise is an interesting development of our work.

Acknowledgments

This work was supported by Tiscali S.p.A., which funded the PhD studies of Ignazio Pillai.

References

A. Androutsopoulos, J. Koutsias, K.V. Cbandrinos and C.D. Spyropoulos. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proceedings of the 23rd ACM International Conference on Research and Developments in Information Retrieval*, pages 160–167, Athens, Greece, 2000.

H.S. Baird and T. Riopka. ScatterType: a reading CAPTCHA resistant to segmentation attack. In *Proceedings of the 12th IS&T/SPIE Conference on Document Recognition & Retrieval*, 2005.

H.S. Baird, M.A. Moll and Sui-Yu Wang. ScatterType: A legible but hard-to-segment CAPTCHA. In: *Proceedings of the eighth International Conference on Document Analysis and Recognition*, 2005.

K. Chellapilla, K. Larson, P.Y. Simard and M. Czerwinski. Building segmentation based human-friendly Human Interactive Proofs (HIPs). In *Proceedings of the Second International Workshop on Human Interactive Proofs*, pages 1–26, 2005.

G.V. Cormack and T.R. Lynam. On-line supervised spam filter evaluation. *ACM Transactions on Information Systems*, forthcoming (available at http://plg.uwaterloo.ca/∼gvcormac/spamcormack.html)

H. Drucker, D. Wu and V.N. Vapnik. Support vector machines for spam categorization. *IEEE Transaction on Neural Networks*, 10(5):1048–1054, 1999.

D. Geer. Will new standards help curb spam? *IEEE Computer*, 47(2):14–16, 2004.

P. Graham. A plan for spam. `http://paulgraham.com/spam.html`, (2002)

N. Holmes. In defense of spam. *IEEE Computer*, 38(4):86–88, 2005

T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges and A. Smola, editors, *Advances in kernel methods - Support vector learning*, pages 41–46, MIT-Press, 1999.

B. Klimt and Y. Yang. The Enron corpus: A new data set for e-mail classification research. In *Proceedings of the European Conference on Machine Learning*, pages 217–226, 2004.

A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *Proceedings of the AAAI Workshop on learning for text categorization*, pages 41–48, 1998.

F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

A. Vinciarelli. Noisy text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1882–1895, 2005.

L. Weinstein. Spam wars. *Communications of the ACM*, 46(8):136, 2003

C.-T. Wu, K.-T. Cheng, Q. Zhu and Yi-L. Wu. Using visual features for anti-spam filtering In *Proceedings of the IEEE International Conference on Image Processing*, Vol. III, pages 501–504, 2005.

M. Sahami, S. Dumais, D. Heckerman and E. Horvitz. A Bayesian approach to filtering junk e-mail. AAAI Technical Report WS-98-05, Madison, Wisconsin, 1998.

X.-L. Wang and I. Cloete. Learning to classify e-mail: A survey. In *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, pages 18–21, 2005.

L. Zhang, J. Zhu and T. Yao. An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing*, 3(4):243–269, 2004.