Improving Image Spam Filtering Using Image Text Features

Battista Biggio, Giorgio Fumera, Ignazio Pillai, Fabio Roli
Electrical and Electronic Eng. Dept., Univ. of Cagliari
Piazza d'Armi, 09123 Cagliari, Italy
{battista.biggio,fumera,pillai,roli}@diee.unica.it

Image-based spam (shortly, image spam) is a trick introduced by spammers few years ago. It consists in embedding all the textual information (i.e., the spam message) into an image attached to the spam e-mail. This allows to evade any filtering module based on the analysis of text in the e-mail's body (usually, a naïve Bayes classifier or a keyword detector). OCR-based modules have been proposed against image spam [7], and simple implementations have been included in spam filters like the popular SpamAssassin.[1] However, besides requiring a relatively high processing time, OCR-based approaches are effective only for clean images, as shown in [4] for SpamAssassin. For this reason, spammers often obfuscate the text embedded into images, making OCR-based approaches ineffective. Other authors proposed to exploit image classification techniques to discriminate between ham and spam images (namely, images attached to ham and spam e-mails), using low-level visual features related for instance to colour distribution, to characteristics of text regions inside an image and to image meta-data [9, 1, 6]. In principle, this approach could be unaffected by text obfuscation techniques used by spammers and has a lower computational cost than OCR-based approaches (especially if classification algorithms as decision trees are used). Classification accuracies between 0.8 and 0.9 were reported in [1, 6] on real and artificial data sets of ham and spam images.

In [8, 2, 3] we proposed a different approach against image spam, complementary to the OCR-based one we proposed in [7]. It was based on the idea that, while clean spam images can be recognised by OCR-based techniques, spam images with obfuscated text could be recognised by the fact that the noise affecting image text is usually of *adversarial* kind, namely it is intentionally added to defeat OCR tools. In other words, when the "signal" (i.e., the spam text) can not be recognised, the presence of adversarial noise (i.e., the consequence of the adversarial action carried out by the spammer to conceal the spam text) could still reveal the image spamminess. To this aim we analysed obfuscation techniques found in real spam images, identified three different kind of image defects common to *different* obfuscation techniques, and devised three measures to detect such kind of image defects and to quantify their amount. Our measures were based on the analysis of connected components in the binary image and of image edges (they are described in detail in [2, 3]). However, the problem of how to integrate such measures in a spam filter architecture (including modules based on OCR tools and possibly on other image classification techniques) was left open.

In subsequent experiments we found that our measures, as defined in [3, 2], detect also some kind of noise present in ham images (due for instance to text placed over a photograph, or to complex characters and backgrounds used in postcards or playbills), leading to too many false positives. This is partly due to the attempt of not focusing these measures to too specific kind of spam noise, to avoid they can be easily evaded (as happened for instance to simple filtering rules based on detecting some typical spam keywords). However, although the pattern of values followed by our measures was different than the desired one (basically, higher values for spam images with obfuscated text and lower values for all other images), we found that it showed some discrimination capability between ham and spam images (even spam images with clean text). This is perhaps due to the fact that the low-level visual characteristics of image text in spam images taken into account by our features (basically, the "shape" of characters in the image text) are different than the ones in ham images. This suggested that our measures could be exploited as additional features to improve the discrimination capability of image classification algorithms between ham and spam images, using the approach proposed in [9, 1, 6]. This was confirmed by the experiments described below.

The experiments were carried out on two corpora of

[1] `http://spamassassin.apache.org`

images (denoted as A and B) attached to real ham and spam personal e-mails. Corpus A was made up of 2,006 ham and 3,297 spam images used in [6]. Corpus B was made up of the same ham images above and of 8,549 spam images collected by the authors.[2] Classification performance was evaluated using five-fold cross-validation, as the false positive (FP) vs the false negative (FN) error rate. We first separately evaluated the performance of the features proposed in [1] and in [6], using the same base classifiers as in the mentioned works (support vector machines (SVM) with radial basis function (RBF) kernel in [1], and C4.5 decision trees in [6]), and of our three features plus the relative text area, which was added to give more weight to low-level image text characteristics found in relatively larger text areas (using a linear SVM classifier). We also evaluated (using a SVM with RBF kernel) another set of "generic" low-level image features devised for the purpose of comparison with the ones in [1, 6]: logarithm of the number of different colours in the image, logarithm of the number of pixels of the image, relative area occupied by the most common colour (used also in [6]), relative area occupied by text (used in [1]). We then evaluated two kind of combinations between our four features and either of the three feature sets mentioned above: combination at the feature level (the two feature vectors are concatenated and a classifier is trained on the resulting feature vector), and combination at the score level (two different classifiers are trained on the two feature set, and the resulting output is combined, using a linear SVM). Results are reported in fig. 1.

It can be seen that features proposed in [1, 6], as well as our "generic" features, outperform our four features when used alone, as expected. However, their discrimination capability is almost always improved when combined (either at feature or at the score level) with our four features, especially for low FP values. We point out that an improvement can be observed also on the corpus A, despite the performance attained by our "generic" features alone was very good (bottom left plot): the FP rate, which was below 0.01 for values of the FN rate higher than 0.05, is reduced to almost 0 when features are combined at the feature level. Moreover, feature level combination seems more effective for small sets of homogeneous features (as for the ones proposed in [1]), while score level combination is better for heterogeneous or large feature sets (as in the case of image meta-data used in [6]).

The general conclusion we can draw is that adding features carefully tailored to *specific* characteristics of spam e-mails (in this particular case, of the text embedded into attached images) can allow to improve the discriminant capability of classification algorithms. This is a topic our research group is currently investigating [5], among others, in the more general context of *adversarial classification* problems (of which spam filtering is a relevant example), with the aim of analysing and developing methods to make classifiers harder to evade. In this context, we also believe that the rationale of the approach proposed in our previous works against image spam with obfuscated text (namely, detecting the presence of embedded text obfuscated in adversarial way) still deserves attention, and could be exploited also for other kind of spam tricks and for other adversarial classification tasks besides spam filtering.

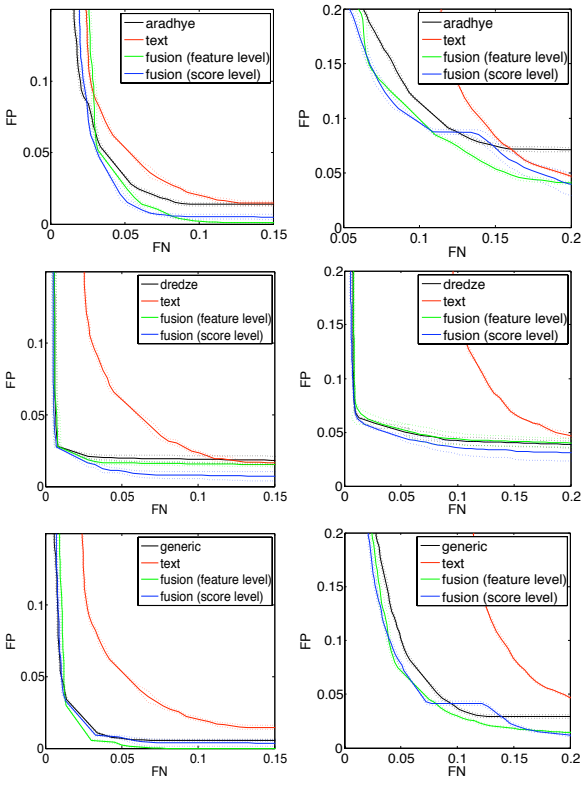

Figure 1: FP vs FN rates (with standard deviation shown as dotted lines) on image corpus A (left) and B (right) obtained by the features proposed in [1] (denoted as 'aradhye', top) and in [6] (denoted as 'drezde', middle), by our "generic" features ('generic', bottom), by our four features ('text', reported in all plots), and by the feature and score level combination between our four features and either of the three other feature sets ('fusion', reported in all plots).

[2]Both corpora are available at `http://prag.diee.unica.it/n3ws1t0/eng/spamRepository`

Acknowledgments

We would like to thank Mark Dredze of Computer and Information Science Dept., University of Pennsylvania, for making his data set publicly available and sending us his code for performing the feature extraction.

We would also like to thank Radhakrishna Achanta and Sabine Süsstrunk of Images and Visual Representation Group (IVRG/LCAV/IC/EPFL, `http://ivrg.epfl.ch/index.html`) for contributing the text detection code[3] for our tests and helping with text detection in about $10,000$ spam images.

References

[1] H. Aradhye, G. Myers, and J. A. Herson. Image analysis for efficient categorization of image-based spam e-mail. In *Proc. Int. Conf. Document Analysis and Recognition*, pages 914–918, 2005.

[2] B. Biggio, G. Fumera, I. Pillai, and F. Roli. Image spam filtering by content obscuring detection. In *Fourth Conference on Email and Anti-Spam (CEAS)*, Microsoft Research Silicon Valley, Mountain View, California, 2-3 August 2007.

[3] B. Biggio, G. Fumera, I. Pillai, and F. Roli. Image spam filtering using visual information. In *14th International Conference on Image Analysis and Processing*, pages 105–110, Modena, Italy, 10-14 September 2007. IEEE Computer Society.

[4] B. Biggio, G. Fumera, I. Pillai, F. Roli, and R. Satta. Evading spamassassin. *Virus Bulletin*, November 2007, http://www.virusbtn.com/vb200711/pdf.

[5] B. Biggio, G. Fumera, and F. Roli. Evade hard multiple classifier systems. In *Workshop on Supervised and Unsupervised Ensemble Methods and their Applications (SUEMA)*, Studies in Computational Intelligence. Springer, 2008. In Press.

[6] M. Dredze, R. Gevaryahu, and A. Elias-Bachrach. Learning fast classifiers for image spam. In *Fourth Conference on Email and Anti-Spam (CEAS)*, Microsoft Research Silicon Valley, Mountain View, California, 2-3 August 2007.

[7] G. Fumera, I. Pillai, and F. Roli. Spam filtering based on the analysis of text information embedded into images. *Journal of Machine Learning Research (special issue on Machine Learning in Computer Security)*, 7:2699–2720, 2006.

[8] G. Fumera, I. Pillai, F. Roli, and B. Biggio. Image spam filtering using textual and visual information. In *MIT Spam Conference*, Cambridge, MA, USA, 30 March 2007.

[9] C.-T. Wu, K.-T. Cheng, Q. Zhu, and Y.-L. Wu. Using visual features for anti-spam filtering. In *Proc. IEEE Int. Conf. on Image Processing*, volume III, pages 501–504, 2005.

[3] `http://lcavwww.epfl.ch/~achanta/TextDetection/TextDetectionResults.html`