On the Study of Anomaly-based Spam Filtering Using Spam as Representation of Normality

Carlos Laorden, Xabier Ugarte-Pedrero, Igor Santos, Borja Sanz, Javier Nieves, Pablo G. Bringas
S^{3}Lab, DeustoTech - Computing
Deusto Institute of Technology, University of Deusto
Avenida de las Universidades 24, 48007, Bilbao, Spain
e-mail: {claorden,xabier.ugarte,isantos,borja.sanz,jnieves,pablo.garcia.bringas}@deusto.es

***Abstract*—In previous work, we presented the first spam filtering method based on anomaly detection that reduces the necessity of labelling spam messages and only employs the representation of legitimate e-mails. This method achieved high accuracy rates detecting spam while maintaining a low false positive rate and reducing the effort produced by labelling spam. In this paper, we study the performance of our previous method when using spam messages to represent normality.**

I. Introduction

Several approaches have been proposed by the academic community to solve the spam problem [1], [2], [3], [4]. Among them, the termed as statistical approaches [5] use machine-learning techniques to classify e-mails. These approaches have proved their efficiency detecting spam and are the most extended technique to fight it. In particular, the use of the Bayes theorem is widely used by the anti-spam filters (e.g., SpamAssasin [6], Bogofilter [7], and Spamprobe [8]).

These statistical approaches are usually supervised, i.e., they need a training set of previously labelled samples. Since these techniques perform better as more training instances are available, a significant amount of previous labelling work is needed to increase the accuracy of the models.

This work includes a gathering phase in which as many e-mails as possible are collected. Then, each e-mail has to be classified as spam or legitimate. Finally, machine-learning models are generated based upon the labelled data. But the availability of labelled training instances is limited and laborious to produce, which slows down the progress of anti-spam systems.

In previous work [9], we proposed the first method that applies anomaly detection to spam filtering. This approach was able to determine whether an e-mail is spam or not by comparing word frequency features with a dataset composed only of legitimate e-mails. If the e-mail under inspection presented a considerable deviation to what it is considered as usual (legitimate e-mails), it can be considered spam. This method reduced the efforts of labelling messages.

The fact that the results were obtained with the minimum distance, which is the most conservative configuration among the distances used in our previous experiment, raised a topic of discussion regarding which e-mails should be considered anomaly e-mails. As said before, for each 5 mail sent worldwide 4 are spam, therefore, in terms of normality, receiving a legitimate e-mail is an anomaly.

In light of this background, we perform in this paper an empirical study of our previous method, only this time, we consider spam e-mails as "normal" messages, and legitimate e-mails as "anomaly". We have used the evaluation method that we employed in our previous work to evaluate this new approach.

II. Common Approaches

Machine-learning approaches model e-mail messages using the *Vector Space Model* (VSM) [10], an algebraic approach for *Information Filtering* (IF), *Information Retrieval* (IR), indexing and ranking. This model represents natural language documents in a mathematical manner through vectors in a multidimensional space.

Formally, let an IR model be defined as a 4-tuple $[\mathcal{E}, \mathcal{Q}, \mathcal{F}, R, (q_i, e_j)]$ [11]: (i) $\mathcal{E}$, is a set of representations of e-mail; $\mathcal{Q}$, is a set of representations of user queries; (ii) $\mathcal{F}$, is a framework for modelling e-mails, queries and their relationships; and (iii) $R(q_i, e_j)$ is a ranking function that associates a real number with a query q_i, $(q_i \in \mathcal{Q})$ and an e-mail representation e_j, $(e_j \in \mathcal{E})$. This function is also called a similarity function.

Let $\mathcal{E}$ be a set of text e-mails e, $\{e : \{t_1, t_2, ...t_n\}\}$, each one comprising an n number of t terms. Consider $w_{i,j}$ a weight for term t_i in an e-mail e_j, whereas if $w_{i,j}$ is not present in e, then $w_{i,j} = 0$. Therefore, an e-mail can be represented as a vector, starting from its origin, of index terms $\vec{e_j} = (w_{1,j}, w_{2,j}, ...w_{n,j})$.

Taking this formalisation as starting point, 'term frequency - inverse document frequency' $(tf - idf)$ [12] is applied to obtain the weight of each word, whereas the weight of the i^{th} word in the j^{th} e-mail, denoted by $weight(i, j)$, is defined by: $weight(i, j) = tf_{i,j} \cdot idf_i$.

The term frequency $tf_{i,j}$ [12] is defined as $tf_{i,j} = m_{i,j} / \sum_k m_{k,j}$, where $m_{i,j}$ is the number of times the word $t_{i,j}$ appears in an e-mail e, and $\sum_k m_{k,j}$ is the total number of word in the e-mail e.

On the other hand, the inverse document frequency idf_i is defined as: $idf_i = |\mathcal{E}|/(|\mathcal{E} : t_i \in e|)$, where $|\mathcal{E}|$ is the total number of documents and $|\mathcal{E} : t_i \in e|$ is the number of documents containing the word $t_{i,j}$.

Then, these approaches apply relevance weights to each feature based on Information Gain (IG) [13], which determines, for each feature, its relevance for the classification of a sample into spam or legitimate e-mail.

Spam filtering techniques use information extracted from e-mails to classify them into 2 categories: spam or legitimate. To this end, most algorithms use VSM, splitting the document into a set of a features (e.g., words, phrases), representing them as vectors, and using these vectors as the basis of the classification.

III. Method Description

Our anomaly detection approach employs the word frequency features of the VSM described to represent e-mails as points in the feature space. In this way, we can obtain a group of e-mails that represent normality (i.e., spam), and decide whether some message is spam or ham by measuring its deviation from the group.

In order to measure the similarity between different e-mails, we compute the following distance measures:

- **Manhattan Distance.** The distance between two points x and y is the sum of the lengths of the projections of the line segment between the two points onto the coordinate axes: $d(x,i) = \sum_{i=0}^{n} |x_i - y_i|$, where x is the first point; y is the second point; and x_i and y_i are the ith component of the first and second point, respectively.
- **Euclidean Distance.** The distance between two points x and y is the length of the line segment connecting v and u. It is calculated as: $d(x,y) = \sum_{i=0}^{n} \sqrt{v_i^2 - u_i^2}$, where x is the first point; y is the second point; and x_i and y_i are the ith component of the first and second point, respectively.

These distances provide a method for measuring the deviation between two e-mails (i.e., the distance between any message and one single message in the group that represents normality). In order to be able to compare a single e-mail against a group of various spam messages, it is necessary to apply a distance selection rule to obtain a unique value dependant on every distance measure performed. To this end, we apply 3 different selection rules: (i) Mean selection rule, which computes the average of the distances to all the members of the spam group; min selection rule, which selects the distance to the nearest spam message; and max selection rule, which returns the distance to the furthest point in the normality representation.

The final deviation value of the e-mail under inspection depends on the distance measure computed and the selection rule applied.

Therefore, when our method inspects an e-mail a final distance value is acquired, which will depend on both the distance measure and the combination metric.

IV. Empirical Validation

We used the *Ling Spam*[1] and *SpamAssassin*[2] datasets.

The SpamAssassin public mail corpus is a selection of 1,897 spam messages and 4,150 legitimate e-mails. Ling Spam consists of a mixture of both spam, 481 e-mails, and legitimate, 2,893 messages, retrieved from the *Linguistic list*, an e-mail distribution list about *linguistics*. From the 4 different datasets provided in this corpus, each one with different pre-process steps, we chose the *Bare* dataset, which had no pre-processing.

We performed for both datasets an *Stop Word Removal* [14] based on an external stop-word list[3] and removed any non alpha-numeric character.

Specifically, we followed the next configuration for the empirical validation:

1) **Cross validation.** For the Ling Spam dataset, we performed a 5-fold cross-validation [15] dividing the dataset of spam e-mails (the normal behaviour) into 5 different divisions of 96 messages, using 4 of them to represent normality and 1 to measure deviations. In this way, each fold is composed of 384 spam e-mails that will be used as representation of normality and 2,508 testing e-mails, from which 96 are spam and 2,412 are legitimate e-mails.
   With regards to the SpamAssassin dataset, we also performed a 5-fold cross-validation [15] dividing the spam e-mails into 4 different divisions of 379 e-mails and 1 division of 380 e-mails, using for of them to represent normality and the other one to test deviations. In this way, each fold is composed of 1,517 or 1,516 spam e-mails that will be used as representation of normality and 4,530 or 4,529 testing e-mails, from which 380 or 379 are spam e-mails and 4,150 are legitimate e-mails. See Table II.

TABLE II
NUMBER OF INSTANCES WITHIN EACH FOLD OF THE 5-FOLD CROSS-VALIDATION PROCESS. NOTE THAT THE NUMBER OF SPAM E-MAILS WITHIN SPAMASSASSIN CORPUS VARIED IN THE FOLDS BECAUSE THE NUMBER OF SPAM E-MAILS WAS NOT DIVISIBLE BY 5.

Ling Spam

	Normality	Deviations	
	# Spam	# Spam	# Legitimate
Fold 1	384	96	2,412
Fold 2	384	96	2,412
Fold 3	384	96	2,412
Fold 4	384	96	2,412
Fold 5	384	96	2,412

SpamAssassin

	Normality	Deviations	
	# Spam	# Spam	# Legitimate
Fold 1	1,516	380	4,150
Fold 2	1,517	379	4,150
Fold 3	1,517	379	4,150
Fold 4	1,517	379	4,150
Fold 5	1,517	379	4,150

2) **Calculating distances and combination rules.** We extracted the aforementioned characteristics and employed the 2 different measures and the 3 different combination rules described in Section III to obtain a final measure

[1] http://nlp.cs.aueb.gr/software_and_datasets/lingspam_public.tar.gz

[2] http://spamassassin.apache.org/publiccorpus/

[3] http://www.webconfs.com/stop-words.php

TABLE I

BEST RESULTS OBTAINED FOR DIFFERENT COMBINATION RULES AND DISTANCE MEASURES. 'THRES'. STANDS FOR THE CHOSEN THRESHOLD.

Ling Spam

Combination	**Manhattan Distance**				**Euclidean Distance**			
	Thres.	**Prec.**	**Rec.**	**F-Meas.**	**Thres.**	**Prec.**	**Rec.**	**F-Meas.**
Mean	8.52995	3.83%	100.00%	7.37%	4.94602	3.83%	100.00%	7.37%
Maximum	9.04422	3.98%	67.92%	7.53%	4.72080	5.23%	35.63%	9.13%
Minimum	0.70134	16.48%	12.50%	14.22%	1.28991	58.73%	15.42%	24.42%

SpamAssassin

Combination	**Manhattan Distance**				**Euclidean Distance**			
	Thres.	**Prec.**	**Rec.**	**F-Meas.**	**Thres.**	**Prec.**	**Rec.**	**F-Meas.**
Mean	5.73362	8.37%	100.00%	15.45%	4.30434	8.37%	100.00%	15.45%
Maximum	8.22796	8.37%	100.00%	15.45%	4.43060	20.75%	56.59%	30.37%
Minimum	0.40784	56.88%	23.10%	32.86%	1.27859	71.58%	40.51%	51.74%

of deviation for each testing evidence. More accurately, we applied Manhattan and Euclidean distances.

For the combination rules we have tested the mean value, the lowest computed distance and the highest computed distance.

3) **Defining thresholds.** For each measure and combination rule, we established 10 different thresholds to determine whether an e-mail is spam or not. These thresholds were selected by first establishing the lowest one. This number was the highest possible value with which no spam messages were misclassified. The highest one was selected as the lowest possible value with which no legitimate spam messages were misclassified.

   In this way, the method is configurable in both reducing false positives or false negatives. It is important to define whether it is better to classify spam as legitimate or to classify legitimate as spam.

4) **Testing the method.** To evaluate the results, we measured the most frequently used in spam filtering: precision (Prec.), recall (Rec.) and f-measure (F-meas.).

   We measured the precision of the spam identification as the number of correctly classified spam e-mails divided by the number of correctly classified spam e-mails and the number of legitimate e-mails misclassified as spam, $Precision = N_{s\to s}/N_{s\to s} + N_{l\to s}$, where $N_{s\to s}$ is the number of correctly classified spam and $N_{l\to s}$ is the number of legitimate e-mails misclassified as spam.

   Additionally, we measured the recall of the spam e-mail messages, which is the number of correctly classified spam e-mails divided by the number of correctly classified spam e-mails and the number of spam e-mails misclassified as legitimate, $Recall = N_{s\to s}/N_{s\to s} + N_{s\to l}$.

   We also computed the F-measure, which is the harmonic mean of both the precision and recall, simplified as follows, $F\text{-}measure = 2N_{s\to s}/2N_{s\to s} + N_{s\to l} + N_{l\to s}$

Table I shows the best results obtained for LingSpam and SpamAssassin using different distances, combination rules and thresholds.

V. DISCUSSION AND CONCLUSIONS

As in our previous work [9] the best results were obtained with the minimum distance, the most conservative configuration for distance. This fact led us to believe that defining legitimate messages as anomaly and spam messages as normal behaviour would offer good results. Therefore, this experiment shows that, at least for the used datasets, using spam messages to represent normality is not a good choice.

REFERENCES

[1] G. Robinson, "A statistical approach to the spam problem," *Linux J.*, vol. 2003, pp. 3–, March 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=636750.636753

[2] P. Chirita, J. Diederich, and W. Nejdl, "MailRank: using ranking for spam detection," in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 373–380.

[3] G. Schryen, "A formal approach towards assessing the effectiveness of anti-spam procedures," in *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, vol. 6. IEEE, 2006, pp. 129–138.

[4] Y. Chiu, C. Chen, B. Jeng, and H. Lin, "An Alliance-Based Anti-spam Approach," in *Natural Computation, 2007. ICNC 2007. Third International Conference on*, vol. 4. IEEE, 2007, pp. 203–207.

[5] L. Zhang, J. Zhu, and T. Yao, "An evaluation of statistical spam filtering techniques," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 3, no. 4, pp. 243–269, 2004.

[6] J. Mason, "Filtering spam with spamassassin," in *HEANet Annual Conference*, 2002.

[7] E. Raymond, "Bogofilter: A fast open source bayesian spam filters," 2005.

[8] B. Burton, "Spamprobe-bayesian spam filtering tweaks," in *Proceedings of the Spam Conference*, 2003.

[9] I. Santos, C. Laorden, X. Ugarte-Pedrero, B. Sanz, and P. G. Bringas, "Anomaly-based spam filtering," in *Proceedings of the 6^{th} International Conference on Security and Cryptography (SECRYPT)*, 2011, in press.

[10] G. Salton, A. Wong, and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[11] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[12] M. McGill and G. Salton, *Introduction to modern information retrieval*. McGraw-Hill, 1983.

[13] J. Kent, "Information gain and a general measure of correlation," *Biometrika*, vol. 70, no. 1, pp. 163–173, 1983.

[14] W. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of information science*, vol. 18, no. 1, pp. 45–55, 1992.

[15] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence*, vol. 14, 1995, pp. 1137–1145.