Int. J. Communications, Network and System Sciences, 2015, 8, 118-129
Published Online May 2015 in SciRes. http://www.scirp.org/journal/ijcns
http://dx.doi.org/10.4236/ijcns.2015.85014

Improving Knowledge Based Spam Detection Methods: The Effect of Malicious Related Features in Imbalance Data Distribution

Ja'far Alqatawna, Hossam Faris, Khalid Jaradat, Malek Al-Zewairi, Omar Adwan

King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan
Email: J.Alqatawna@ju.edu.jo

Received 1 April 2015; accepted 29 April 2015; published 30 April 2015

Abstract

Spam is no longer just commercial unsolicited email messages that waste our time, it consumes network traffic and mail servers' storage. Furthermore, spam has become a major component of several attack vectors including attacks such as phishing, cross-site scripting, cross-site request forgery and malware infection. Statistics show that the amount of spam containing malicious contents increased compared to the one advertising legitimate products and services. In this paper, the issue of spam detection is investigated with the aim to develop an efficient method to identify spam email based on the analysis of the content of email messages. We identify a set of features that have a considerable number of malicious related features. Our goal is to study the effect of these features in helping the classical classifiers in identifying spam emails. To make the problem more challenging, we developed spam classification models based on imbalanced data where spam emails form the rare class with only 16.5% of the total emails. Different metrics were utilized in the evaluation of the developed models. Results show noticeable improvement of spam classification models when trained by dataset that includes malicious related features.

Keywords

Spam E-Mail, Malicious Spam, Spam Detection, Spam Features, Security Mechanism, Data Mining

1. Introduction

In the context of the Internet and the World Wide Web, spam is a general term which is usually used to refer to

How to cite this paper: Alqatawna, J., *et al*. (2015) Improving Knowledge Based Spam Detection Methods: The Effect of Malicious Related Features in Imbalance Data Distribution. *Int. J. Communications, Network and System Sciences*, **8**, 118-129. http://dx.doi.org/10.4236/ijcns.2015.85014

unsolicited commercial communications most likely in a form of email messages [1]-[3]. While our study concerns with emails spam, other spam forms do exist including social network spam, blog spam, forum spam and search engine spam [4]. Spam email is used for advertising products and services typically related to adult entertainment, quick money and other attractive merchandises [5]. Most of the time, customers are unwilling to receive such form of communication as it is without customers' consent to receive its related ads or even to opt out of that. Among early spammers' activities, which can be dated back to the mid of 1990s, is the development of the first commercial spamware "Flood Gage" which gave the ability to collect emails from various sources and then to send thousands of emails per hour [2]. Stone-Gross, *et al*. have analyzed the underground economy of spam and estimated that in a single affiliate program, spammer revenue can exceed one million dollars per month [6].

The large volume of spam traffic negatively affects networks bandwidth, email servers storage and processing power, user time and work productivity [7]. According to Kaspersky's security bulletin, during 2013 spam counted for ~70% of the total email traffic [8]. Notability, the bulletin showed that there was a shift toward criminalizing commercial spam. The amount of spam containing malicious contents increased compared to the one advertising legitimate products and services. Such situation adds another dimension to the adverse nature of spam email, which mainly touches on the privacy and security of individuals and organizations.

Accurate spam detection is considered a difficult task. According to [9], this is partially due to several reasons including the subjective nature of spam, concept drift, language issues, processing overhead and message delay, and the irregular cost of filtering errors. Spam contents exhibit a high subjective nature; for instance, a message containing several drug names might be a spam, but it might not be the case if the message is exchanged in a context of medical organizations. Concept drift is another reason that affects filtering accuracy because spammers could generate contents in unpredictable patterns. For instance, the word offer can be written as O_F_F_E_R or 0ff3R. Additionally, detection methods rarely consider the different languages that might be used to compose email messages. Moreover, checking for spam adds extra processing time and delay to the delivery of email messages. Furthermore, the acceptance of classification errors varies from user to use which affect the usability of the various available solutions.

In this paper, we propose a framework to improve knowledge based spam detection methods. The issue of spam detection is investigated with the aim to develop an efficient method to identify spam email based on the analysis of the content of email messages. While considerable volume of spam email is in fact a malicious one, combining the features extracted from the malicious spam content with the typical spam features significantly decreases the error rates of the classical spam classifiers. To make the problem more challenging, we developed spam classification models based on imbalanced data where the spam class forms the rare class with only 16.5% of the total emails. Different metrics were utilized in the evaluation and comparison of the developed models. Results show noticeable improvement of spam classification models when trained by datasets that includes malicious related features. The rest of the paper is organized as follows: in the next section, we address the email spam issue in the literature. In Section 3, we present our proposed framework for spam detection and discuss the results in Section 5. We conclude our findings in Section 6 and provide a comprehensive list of the features employed by our framework in Appendix A.

2. Background

2.1. Spam Detection Methods

Apart from legislations, which act as a deterrence control, one possible way to detect and prevent spam is to apply a rule-based filter [10] [11]. Such filtering approach can be applied to the header of the message to check that the source address does not belong to a spammer domain. Additionally, it can be applied to the content of the message to check for the existence of text patterns and words that usually used by spammers. In order for this method to be affective, a rather large number of rules are needed. On the other hand, spammers could bypass a rule-based filter by forging the source of email and/or obfuscating the contents that might help the filter to classify incoming emails as spam [3] [11].

Another common approach for spam detection is realized using learning-based filter. The basic idea of this approach is to train the filter in order to extract the knowledge that can be used to detect spam. Such training uses a large dataset containing spam emails along with legitimate ones, and then the filter can use the extracted knowledge to classify new emails. Most of the techniques under this filtering approach utilize Machines

Learning algorithms such as Naive Bayes Classifier, Support Vector Machines and Artificial Neural Networks [1].

Due to the dynamic nature of both content and structure of spam emails, it is argued that one of the serious problems with the common learning-based classifiers used for spam detection is the lack of incremental learning capability [7]. This means that the filter needs to be retrained with the updated dataset to accommodate changes in the new messages. The dramatically increase in the email spam and the adaptive nature of spam generator call for investigating a more efficient and adaptive approach for spam detection [12]. Such approaches should be designed to allow dynamic addition or removal of feature without re-building the entire filter [1].

2.2. Spam Features

Spam detection is based on the assumption that its content differs from that of a legitimate email in ways that can be quantified. Caruana & Li argued that spam emails have a number of similar characteristics in terms of structure, content and diffusion approaches [13]. A content of a typical spam email, which advertises attractive products and services can be characterized by several statically features such as particular words frequency, special characters frequency, digits and/or alphabet frequency. For instance, Symantec security survey reported that emails containing sexual and dating related contents represented 55% of all spam traffic in 2012 [14]. In a study conducted in [15], 100 emails have been analyzed and spammer typical words (e.g. *winner*, *dollar*, *award*, *cash prize*, *top job opportunities*, *earn more*, *beneficiary*, *good news*, *claim*, *high salary and payment*) were extracted and ranked. Such words attract users and increase the chance that spam will be opened. Luckner, Gad, & Sobkowiak suggested that spam might contain "weird combinations" which mean the existence obfuscated words represented by a string of lower case letters with some upper-case letters or digits among elements of the string such as Credit4U, v1agra and StaffForFree [16]. Lee *et al.* has applied parameters optimization and feature selection on 57 features used in the Spambase dataset [17]. Their study showed that using only 19 features it was possibly to classify spam email with a 95.0011 detection rate. **Table 1** shows the top 10 features used in their study.

2.3. Malicious Spam Features

Malicious email content represents another dimension for the spam issue. A report released by Kaspersky showed that 3.2% of the email traffic was carrying malicious attachments such as Trojans, Worms and Spyware [8]. The top of these distributed malware was a Trojan masquerading as Web registration page of an online banking service utilized in a distributed phishing campaign to steal users' credential. As a deception technique, malicious spammers usually change the file extension to disguise malicious attachments. Extension such as .zip, .rar, .pdf, and .jpg are usually used [18]. Symantec reported that 23% of spam traffic in 2012 contained URLs that pointed to malicious websites [14]. The report revealed several aspects, which characterized these malicious URLs including the use of nest of URL shortening services and HTTPS to trick users to trust these links.

Table 1. Spam base data set top 10 features [17].

Rank	Feature	Average Variable Importance
1	char_freq_!	0.5021
2	word_freq_remove	0.4838
3	word_feq_credit	0.4740
4	char_feq_$	0.4739
5	word_feq_hp	0.4725
6	word_feq_edu	0.4687
7	capital_run_length_longest	0.4644
8	word_feq_free	0.4490
9	capital_run_length_total	0.4448
10	word_feq_george	0.4431

Alazab & Broadhurst emphasized the role of spam emails with malicious attachments and URLs as a source of malware infection. They studied three real world datasets containing over 13 million spam emails collected in 2012 by the Australian Internet Security Initiative (AISI) where spam emails were harvested from several sources that utilize different techniques to detect spam emails mainly user labeling, machine learning, spam-traps, and commercial spam filters. The three datasets come from the spam reporting add-on "HabuL", which is used by a popular email client (Thunderbird), a global system of honeypots and spam-traps and the spam filter from an Australian ISP. Attachments and URLs were submitted to VirusTotal to identify spam emails with malicious content. Interestingly, 21.4% and 22.3% had at least one attachment and URL with malicious content. Their results showed that 90% of the malicious files were in fact compressed (.zip) files and that multiple obfuscation technique were deployed in order to disguise the true file extension such as the using of double extension, long file name, fake icons and the Right-to-Left Override (RLO) special Unicode character (U + 202E). Moreover, both URL shortening services and daisy chained shortened links were used to conceal the true destination of malicious URLs and avoid detection [18]. The obfuscation techniques identified in their study can serve as features to detect spam emails with malicious contents.

Tran, Alazab, & Broadhurst have applied several vandalism detection features in spam detection to identify spam emails with malicious contents in addition to defining a new set of features. A combination of 58 features were classified into five main categories relevant to where in the email the feature resides, which are header features, subject features, payload (body) features, attachments and URLs features, where the latter two features classifications are used independently from each other [19]. A random forest classifier was used in order to rank the features across both the "HabuL" and spam-traps (Botnet) datasets used in [18] based on their entropy. **Tables 2(a)-(d)** presents the top 5 features based on the spam emails from the month of November.

Le Blond *et al.* studied targeted social engineering attacks that employ emails with malicious attachments using a data set consists of 1493 emails with 1176 malicious attachments collected over a four-year period (2009-2013) by two members of the World Uyghur Congress (a Chinese human-rights non-governmental organization). Their results indicate that the language, topic, and timing of the emails were highly tailored to the recipients. They also showed that the attackers have masqueraded the sender address with several techniques

Table 2. Top 5 spam detection features [19].

Spam Attachments Features					
Habul Dataset			Botnet Dataset		
Rank	Category	Feature	Rank	Category	Feature
1	Subject	Number of capitalized words	1	Subject	Min of the compression ratio for the bz2 compressor
2	Subject	Sum of all the character lengths of words	2	Subject	Min of the compression ratio for the zlib compressor
3	Subject	Number of words containing letters and numbers	3	Subject	Min of character diversity of each word
4	Subject	Max of ratio of digit characters to all characters of each word	4	Subject	Min of the compression ratio for the lzw compressor
5	Header	Hour of day when email was sent	5	Subject	Max of the character lengths of words
		(a)			(b)
Spam URLs Features					
1	URL	The number of all URLs in an email	1	Header	Day of week when email was sent
2	URL	The number of unique URLs in an email	2	Payload	Number of characters
3	Payload	Number of words containing letters and numbers	3	Payload	Sum of all the character lengths of words
4	Payload	Min of the compression ratio for the bz2 compressor	4	Header	Minute of hour when email was sent
5	Payload	Number of words containing only letters	5	Header	Hour of day when email was sent
		(c)			(d)

such as email address spoofing and using email addresses familiar to the receiver but with minor hard-to-notice differences, which accounted for 30% and 41% of the emails respectively. The malicious attachments were analyzed through Virus Total and dynamic taint analysis [20].

Amin defined 65 different features to detect targeted emails with malicious contents. These features are categorized into two broad categories; persistent threat and recipient oriented features. Persistent threat features are tightly coupled with the attacker's environment such as IP address, time zone, character encoding, and tools. On the other hand, recipient oriented features are related to the spam victim; such as his/her role in an organization, the relationship between a person and another entity, and the level of access he/she has [21].

3. A Spam Detection Framework

Our work to improve spam detection methods is based on the assumption that a considerable volume of spam emails is in fact a malicious one. Combining the features extracted from the malicious spam content with the typical spam features could significantly decrease the error rates of the classical spam classifiers. Based on the analysis presented in Sections 2.2 and 2.3 we were able to specify a set of features that we have used to develop a rich model to characterize spam emails taken into account the features of malicious contents, which could better improve the detection of spam emails. As shown in **Table 3**, 90 features were used in this study. Appendix A lists a complete description of the proposed features. 60 out of the 90 features are identified as malicious related features based on the literature review of the domain.

In order to study the effectiveness of these features in building accurate spam prediction models, we propose the framework shown in **Figure 1**. First, we developed a Java software to extract the aforementioned features from the header, attachment and body parts of each email. Spam Assassin public mail corpus served as our data set, which consists of (5051) ham emails and (1000) spam emails [22]. Based on the extracted features, we develop four different spam detection models using classical data mining classification techniques. These techniques include the following four classification algorithms:

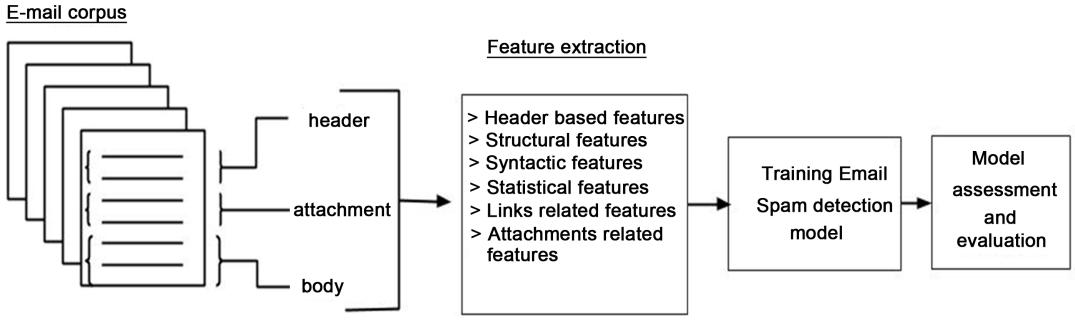


Figure 1. The proposed spam detection framework.

Table 3. Proposed spam detection features.

Category	Email Part	No. of Features
Header based	Header/Subject	31
Character based	Body	7
Word based	Body	14
Syntactic features	Body	10
Structural features	Body	9
Particular content (world/character)	Body	10
Size	Body	1
Links related features	Body	2
Attachment related features	Attachment	6

- The C4.5 algorithm which is one of the most famous classification algorithms used for generating decision trees. C4.5 is based on the concept of information entropy.
- The Multilayer Perceptron Neural Network (MLP), which is a mathematical model and a type of feed forward neural network. The basic component of the MLP is "neuron" which is a simple processing element. Neurons are arranged in layers and each layer is fully connected with the next one by means of weights. MLP is trained by updating these weights until a predetermined level of error is reached. In this work we used the back propagation algorithm for training the MLP.
- Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayesian theorem with independence assumptions between the features. The main advantage of this classifier is its scalability and that it could perform well with large datasets.
- Random forest algorithm which is a powerful ensemble algorithm and mainly depends on decision tree classifiers. Random forest generates and combines a predetermined number of decision trees at training time where each tree depends on the values of a random vector sampled independently. A Random forest classifier gives a classification output by taking the majority vote of its decision trees.

Finally, the developed models were evaluated using different evaluation ratios which will e discuss the following section.

4. Experiments and Results

4.1. Experiments Setup

In order to evaluate the effectiveness of the malicious related features in helping the classification algorithms in detecting the spam emails, we developed and tested standard classification algorithms based on two groups of features. First, we used all the features described in Section 3, which included the malicious related features, while the second group contains all features except malicious related features. In our experiments, we use C4.5 decision tree classifier, Multilayer Perceptron Neural Networks (MLP), Naïve Bayes classifier and Random Forests. For Random Forests, number of trees is set empirically to 10, while the parameters of the MLP are tuned as shown in **Table 4**. Finally in order to obtain reliable results, we apply 5 folds cross validation for training and testing.

4.2. Evaluation Measurements

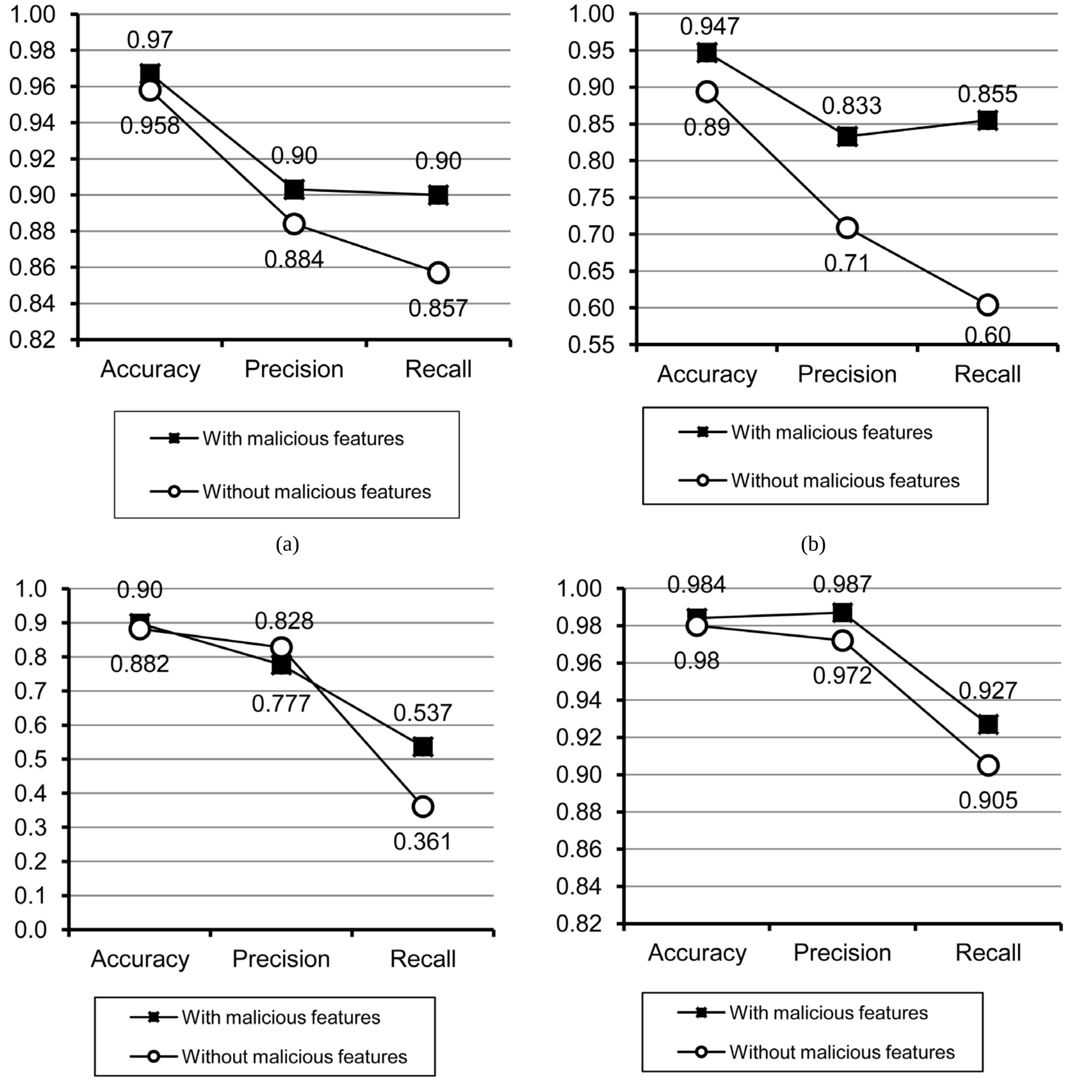
The developed spam classification models in this work are evaluated by referring to the confusion matrix shown in **Table 5**. The confusion matrix shows the four possible results which are: The email is Ham and the classifier predicted it correctly so it is True Ham (TH); The email is spam and the classifier predicted it correctly so it is True Spam (TS) and the other two wrong possibilities which are the email is Ham and the classifier predicted it wrongly so it is False Spam (FS) and finally the email is Spam and the classifier predicted it as Ham so it is False Ham (FH). Based on this confusion matrix we calculate Accuracy rate, Precision and Recall as shown in Equations (1)-(3) respectively. Accuracy measures the rate of correctly classified instances of both classes: spam and ham. Since our data set has imbalance data distribution, where the number of spam emails is much smaller than the number of ham emails, Precision and Recall rates are calculated for the spam class. This provides more consideration to the spam class, which forms the rare class in our case.

4.3. Results and Discussion

The results for evaluating the four classifiers, *i.e.* C4.5, MLP, Naïve Bayer and Random Forest, are shown in **Figures 2(a)-(d)** respectively. In each figure, we show the values of the accuracy, precision and recall for both

Table 4. MLP classifier tuning settings.

Parameter	Value
Epochs	2000
Learning rate	0.3
Momentum	0.2
Activation function	$\frac{1}{1+e^{-x}}$



(a) (b)

(c) (d)

Figure 2. The evaluation results of the four spam classifiers. (a) Evaluation results of C4.5 classifier; (b) Evaluation results of MLP classifier; (c) Evaluation results of Naive Bayes classifier; (d) Evaluation results of Random Forest classifier.

Table 5. Confusion matrix.

	Predicted	
	Ham	Spam
Actual Ham	True Ham (TH)	False Spam (FS)
Actual Spam	False Ham (FH)	True Spam (TS)

$$\text{Accuracy} = \frac{(\text{TH} + \text{TS})}{\text{TH} + \text{FS} + \text{FH} + \text{TS}} \tag{1}$$

$$\text{Precision} = \frac{TS}{TS + FS} \tag{2}$$

$$\text{Recall} = \frac{TS}{TS + FH} \tag{3}$$

sets of features (with and without malicious related features). Notable our proposed work to add the malicious related features to the training data has improved the accuracy rate for all classifiers. While on the other hand, those features significantly improved the precision and recall values for the small class, which represents in our case the spam emails class. Recall rates for C4.5, MLP, Naïve Bayer and Random Forest were also improved by 4%, 25%, 17% and 2% respectively. Precision rates for C4.5, MLP and Random Forest were enhanced by 2%, 13% and 1.5% respectively. Our results showed an overall improvement over all the evaluation measures for all four classifiers by adding malicious related features with the exception of the precision ratio for Naïve Bayes, which was slightly decreased. We can conclude that although the problem is challenging for the classical classifiers due to the imbalance data distribution of the data collected, adding malicious spam related features could improve the learning process of these classifiers in detecting the rare class, which is the emails spam class.

5. Conclusion

In this paper, we investigated the effect of adding considerable number of malicious related features to the data used for training classical data mining classifiers. Four classifiers including C4.5 decision trees, MLP, Naïve Bayes and Random forests were applied. In order to make the problem more challenging, we trained and tested those classifiers on imbalanced dataset were the spam class forms only 16.5% of the whole data. Evaluation results of the developed classification models showed that adding the malicious related features has significantly improved the ability of the classifiers to detect spam emails.

6. Future Work

As this work showed the importance of using features related to malicious emails in identifying the spam email in general, more investigation is planned to be done regarding these features. For example, more possible features related to malicious emails could be included in the knowledge base that is used for training spam identification systems.

References

[1] Guzella, T.S. and Caminhas, W.M. (2009) A Review of Machine Learning Approaches to Spam Filtering. *Expert Systems with Applications*, **36**, 10206-10222. http://dx.doi.org/10.1016/j.eswa.2009.02.037

[2] Rao, J.M. and Reiley, D.H. (2012) The Economics of Spam. *Journal of Economic Perspectives*, **26**, 87-110. http://dx.doi.org/10.1257/jep.26.3.87

[3] Stern, H. and Others (2008) A Survey of Modern Spam Tools. 5*th Conference on Email and Anti-Spam*, CEAS, California.

[4] Kanich, C., Weaver, N., McCoy, D., Halvorson, T., Kreibich, C., Levchenko, K., Paxson, V., Voelker, G.M. and Savage, S. (2011) Show Me the Money: Characterizing Spam-Advertised Revenue. *USENIX Security Symposium*, San Francisco, August 2011, 15.

[5] Cranor, L.F. and LaMacchia, B.A. (1998) Spam! *Communications of the ACM*, **41**, 74-83. http://dx.doi.org/10.1145/280324.280336

[6] Stone-Gross, B., Holz, T., Stringhini, G. and Vigna, G. (2011) The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-Scale Spam Campaigns. *USENIX Workshop on Large-Scale Exploits and Emergent Threats* (*LEET*), Boston, March 2011.

[7] Su, M.-C., Lo, H.-H. and Hsu, F.-H. (2010) A Neural Tree and Its Application to Spam E-Mail Detection. *Expert Systems with Applications*, **37**, 7976-7985. http://dx.doi.org/10.1016/j.eswa.2010.04.038

[8] Gudkova, D. (2013) Kaspersky Security Bulletin. Spam Evolution 2013.

[9] Pérez-Díaz, N., Ruano-Ordás, D., Fdez-Riverola, F. and Méndez, J.R. (2012) SDAI: An Integral Evaluation Methodology for Content-Based Spam Filtering Models. *Expert Systems with Applications*, **39**, 12487-12500. http://dx.doi.org/10.1016/j.eswa.2012.04.064

[10] Kamboj, R. (2010) A Rule Based Approach for Spam Detection. Thapar University, Patiala.

[11] Pérez-Díaz, N., Ruano-Ordás, D., Méndez, J.R., Gálvez, J.F. and Fdez-Riverola, F. (2012) Rough Sets for Spam Filtering: Selecting Appropriate Decision Rules for Boundary E-Mail Classification. *Applied Soft Computing*, **12**, 3671-

3682. http://dx.doi.org/10.1016/j.asoc.2012.05.024

[12] Idris, I., Selamat, A., Thanh Nguyen, N., Omatu, S., Krejcar, O., Kuca, K. and Penhaker, M. (2015) A Combined Negative Selection Algorithm-Particle Swarm Optimization for an Email Spam Detection System. *Engineering Applications of Artificial Intelligence*, **39**, 33-44. http://dx.doi.org/10.1016/j.engappai.2014.11.001

[13] Caruana, G. and Li, M. (2012) A Survey of Emerging Approaches to Spam Filtering. *ACM Computing Surveys* (*CSUR*), **44**, 9. http://dx.doi.org/10.1145/2089125.2089129

[14] Symantec (2013) Internet Security Threat Report 2013.

[15] Santhi, G., Wenisch, S.M. and Sengutuvan, P. (2013) A Content Based Classification of Spam Mails with Fuzzy Word Ranking. *IJCSI International Journal of Computer Science Issues*, **10**, 48-58.

[16] Luckner, M., Gad, M. and Sobkowiak, P. (2014) Stable Web Spam Detection Using Features Based on Lexical Items. *Computers & Security*, **46**, 79-93. http://dx.doi.org/10.1016/j.cose.2014.07.006

[17] Lee, S.M., Kim, D.S., Kim, J.H. and Park, J.S. (2010) Spam Detection Using Feature Selection and Parameters Optimization. *International Conference on Complex*, *Intelligent and Software Intensive Systems* (*CISIS*), Poland, February 2010, 883-888.

[18] Alazab, M. and Broadhurst, R. (2014) Spam and Criminal Activity. *Trends and Issues* (*Australian Institute of Criminology*), Forthcoming. http://dx.doi.org/10.2139/ssrn.2467423

[19] Tran, K.-N., Alazab, M. and Broadhurst, R. (2013) Towards a Feature Rich Model for Predicting Spam Emails Containing Malicious Attachments and URLs. 11*th Australasian Data Mining Conference*, Canberra, November 2013.

[20] Le Blond, S., Uritesc, A., Gilbert, C., Chua, Z.L., Saxena, P. and Kirda, E. (2014) A Look at Targeted Attacks through the Lense of an NGO. *Proceedings of the* 23*rd USENIX Conference on Security Symposium*, San Diego, August 2014, 543-558.

[21] Amin, R.M. (2011) Detecting Targeted Malicious Email through Supervised Classification of Persistent Threat and Recipient Oriented Features. The George Washington University, Washington DC.

[22] Spam Assassin Project (2015) Spam Assassin Public Corpus. https://spamassassin.apache.org/publiccorpus/

Appendix A

Category	Description	Type
Subject	Year	Malicious
Subject	Month of year when email was sent	Malicious
Subject	Day of week when email was sent	Malicious
Subject	Hour of day when email was sent	Malicious
Subject	Minute of hour when email was sent	Malicious
Subject	Second of Minute when email was sent	Malicious
Subject	From header email address domain is google.com	Malicious
Subject	From header email address domain is aol.com	Malicious
Subject	From header email address domain is .gov	Malicious
Subject	From header email address domain is hotmail.com	Malicious
Subject	From header email address domain is .mil	Malicious
Subject	From header email address domain is yahoo.com	Malicious
Subject	From header email contains example.com	Malicious
Subject	Received line contains redacted	Malicious
Subject	Message id contains redacted	Malicious
Subject	Replay to header is defined	Malicious
Subject	Replay to different then from	Malicious
Subject	Replay to email address is at gmail.com	Malicious
Subject	Replay to email address is at hotmail.com	Malicious
Subject	Replay to email address is at yahoo.com	Malicious
Subject	Replay to email address is at aol.com	Malicious
Subject	Replay to email address is at .gov	Malicious
Subject	Replay to email address is at .mil	Malicious
Subject	To header is defined but empty	Malicious
Subject	To email address is at gmail.com	Malicious
Subject	To email address is at hotmail.com	Malicious
Subject	To email address is at yahoo.com	Malicious
Subject	To email address is at aol.com	Malicious
Subject	To email address is at .gov	Malicious
Subject	To email address is at .mil	Malicious
Subject	To MSN	Malicious
Subject	To localhost	Malicious
Subject	To email address is at “example.com”	Malicious
Subject	Forward-to header is defined	Malicious
Subject	X-Mailer-Version	Malicious

Continued

Email Body: Character	Total number of digit character	Statistical
Email Body: Character	Total number of white space	Statistical
Email Body: Character	Total number of upper case character	Statistical
Email Body: Character	Total number of characters	Statistical
Email Body: Character	Total number of tabs	Statistical
Email Body: Character	Total number of special characters	Statistical
Email Body: Character	Total number of alpha characters	Statistical
Email Body: Word	Total number of words	Statistical
Email Body: Word	Average word length	Statistical
Email Body: Word	Vocabulary richness	Statistical
Email Body: Word	Words longer than 6 characters	Statistical
Email Body: Word	Total number of words (1 - 3 Characters)	Statistical
Email Body: Word	Entropy measure	Malicious
Email Body: Word	Hapax legomena	Malicious
Email Body: Word	Hapax dislegomena	Malicious
Email Body: Word	Yule's K	Malicious
Email Body: Word	Sichles S	Malicious
Email Body: Word	Honores R	Malicious
Email Body: Word	Word length frequency distribution	Statistical
Email Body: Syntactic	Number of single quotes	Statistical
Email Body: Syntactic	Number of commas	Statistical
Email Body: Syntactic	Number of periods	Statistical
Email Body: Syntactic	Number of semi-colons	Statistical
Email Body: Syntactic	Number of question marks	Statistical
Email Body: Syntactic	Number of multiple question marks	Statistical
Email Body: Syntactic	Number of exclamation marks	Statistical
Email Body: Syntactic	Number of multiple exclamation marks	Statistical
Email Body: Syntactic	Number of colons	Statistical
Email Body: Syntactic	Number of ellipsis	Statistical
Email Body: Structural	Total Number of lines	Statistical
Email Body: Structural	Total number of sentences	Statistical
Email Body: Structural	Total number of paragraphs	Statistical
Email Body: Structural	Average number of sentences per paragraph	Statistical
Email Body: Structural	Average number of words pre paragraph	Statistical
Email Body: Structural	Average number of character per paragraph	Statistical
Email Body: Structural	Average number of word per sentences	Statistical

Continued

Email Body: Structural	Number of sentence begin with upper case	Statistical
Email Body: Structural	Number of sentence begin with lower case	Statistical
Email Body: Paper	Character frequency "!"	Malicious
Email Body: Paper	Character frequency "$"	Malicious
Email Body: Paper	Word frequency remove	Malicious
Email Body: Paper	Word frequency credit	Malicious
Email Body: Paper	Word frequency hp	Malicious
Email Body: Paper	Word frequency edu	Malicious
Email Body: Paper	Word frequency free	Malicious
Email Body: Paper	Word frequency George	Malicious
Email Body: Paper	Capital run length longest	Malicious
Email Body: Paper	Capital run length total	Malicious
Email Body: General	Email size	Malicious
Attachments: General	No. of URL	Malicious
Attachments: General	URL unique	Malicious
Attachments: General	Multipart/Mixed	Malicious
Attachments: General	Multipart/Mixed unique	Malicious
Attachments: General	Multipart/Alternative	Malicious
Attachments: General	Multipart/Alternative unique	Malicious
Attachments: General	Text/Plain	Malicious
Attachments: General	Text/Plain unique	Malicious