Latent Semantic Analysis and Keyword Extraction for Phishing Classification

Gastón L'Huillier, Alejandro Hevia
Department of Computer Science
University of Chile
Blanco Encalada 2120, Santiago, Chile
Email: {glhuilli,ahevia}@dcc.uchile.cl

Richard Weber, Sebastián Ríos
Department of Industrial Engineering
University of Chile
República 701, Santiago, Chile
Email: {rweber,srios}@dii.uchile.cl

***Abstract*—Phishing email fraud has been considered as one of the main cyber-threats over the last years. Its development has been closely related to social engineering techniques, where different fraud strategies are used to deceit a naïve email user. In this work, a latent semantic analysis and text mining methodology is proposed for the characterisation of such strategies, and further classification using supervised learning algorithms. Results obtained showed that the feature set obtained in this work is competitive against previous phishing feature extraction methodologies, achieving promising results over different benchmark machine learning classification techniques.**

***Index Terms*—Phishing detection, Text mining, Latent Semantic Analysis**

I. Introduction and Previous Work

Nowadays, in the *cyber-crime* context, one of the most common social engineering threats is the phishing fraud. This malicious activity consists of sending email scams, asking for personal information to break into any site where victims may store useful private information, such as financial institutions, e-commerce and other web sites. By this methods, millions of dollars are stolen every year, and this number is likely to keep raising as the internet penetration in our everyday life increases.

Different text mining techniques for phishing filtering have been proposed. In [1], Logistic Regression, Support Vector Machines (SVMs), and Random Forests are used to estimate classifiers for the correct labeling of email messages. By using of more sophisticated text mining techniques, Bergholz et al. ([3], [4]) proposed a novel characterization of emails using a Class-Topic model. For phishing feature extraction several methodologies have been developed [1], [2], [4], [7], while for phishing classification data mining approaches have been used [7], [8].

The main contribution of this work is a feature extraction methodology for phishing emails that, using latent semantic analysis features and keyword extraction techniques, enhances traditional machine learning algorithms used in email filering (such as Support Vector Machines, naïve Bayes, and logistic regression).

This paper is structure as follows: In section II, the proposed feature extraction and selection methodology is presented. The experiments and results are presented in section III, and final conclusions in section IV.

II. Latent Semantic Analysis and Keyword Extraction for Phishing Features

In this paper the applicability of topic based features for malicious message filtering is determined by text mining methodologies. The proposed characterization methodology of email messages is described in figure 1.

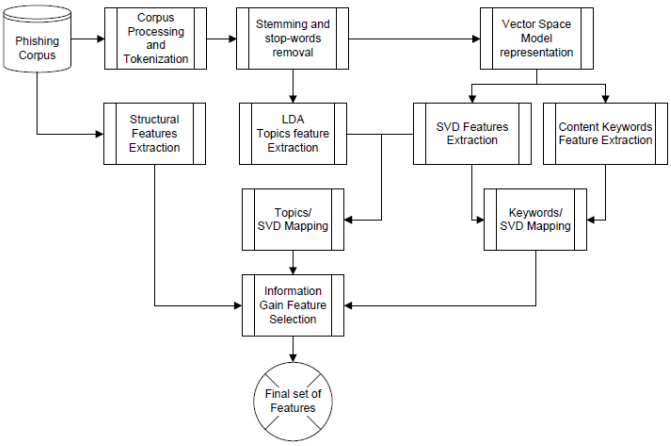

Fig. 1. Feature extraction flow diagram process for phishing messages.

Let the set of features determined by the keyword finding algorithm be Ω, the set of features determined by a Singular Value Decomposition (SVD) of the Vector Space Model (VSM) representation of the corpus be Υ and the set of features determined by Latent Dirichlet Allocation (LDA) Γ, and the set of basic structural features be Ξ, then the final set of features Π that is analysed into the feature extraction step is given by,

$$\Pi = \Xi \cup ((\Gamma \cap \Upsilon) \cup (\Omega \cap \Upsilon)) \quad (1)$$

As shown in equation 1, the final set of features Π is a combination of structural basic features Ξ, which are independent from the other content based features set Γ, Ω and Υ. However, these sets are not independent from each other. They are represented by binary features, indicating whether a keyword or topic is presented in a given message, whose intersection describes a final set of features that represents a training set of phishing messages.

A. Keyword Extraction

As proposed in [11], a keyword extraction proceedure has been used in different web content and web usage mining applications. The idea is to extract words that represents a significant meaning from a given set of documents. This method is based on clustering techniques over the VSM for a given corpus. In our work, a k-means clustering with the cosine similarity distance between documents VSM was used.

B. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [5] is a model where latent topics of documents are infered from estimated probability distributions over the training dataset. The key idea behind LDA, is that every topic is modeled as a probability distribution over the set of words represented by the vocabulary, and every document as a probability distribution over a set of topics. By using this text-mining method, different topics can be extracted and used as input features for the phishing classification task.

C. Singular Value Decomposition

As described in [9], SVD preserves the relative distances in the VSM matrix, while projecting it into a Semantic Space Model (SSM), which has a lower dimensionality. This allows to keep just the minimum information needed to define the appropiate representation of the dataset. Furthermore, the SVD of the *tf-idf* matrix reveals the underlying semantic relationship between terms and documents.

III. Experiments and Results

The classification of phishing emails is a natural extension of text mining, where the most promising classification algorithms are Support Vector Machines, naïve Bayes, Random Forest, among other text categorization algorithms [10]. Below, the experimental setup, evaluation criteria, and results obtained are presented.

A. Experimental Setup and Evaluation Criteria

A 10×10 cross-validation learning schema using benchmark machine learning algorithms on the complete database characterized with different set of features was developed. The learning algorithms were implemented using open source machine learning algorithms: SVMs were constructed using the libSVM-library [6]; the naïve Bayes model, and the logistic regression method were implemented in Weka [12].

Feature sets used as benchmark were evaluated over the F-Measure performance criteria. The list of benchmark sets used is presented:

1) Structural features[1] represented by the feature set (Ξ)
2) SVD features represented by the feature set (Υ)
3) Content-topic features represented by the feature set (Γ)
4) Keywords features represented by the feature set (Ω)
5) SVD, content-topic and keyword features intercepted, represented by the feature set, $\hat{\mathcal{F}} = (\Omega \cap \Upsilon) \cup (\Gamma \cap \Upsilon)$.
6) All features, considering $\mathcal{F} = \hat{\mathcal{F}} \cup \Xi$
7) All features $\mathcal{F}$ preprocessed by the mutual information feature selection algorithm.

All feature sets were evaluated using SVMs, naïve Bayes, and logistic regression classification algorithms. Also, results for SVMs were compared with those obtained by [4] and [7] whose experimental setup is based on the same corpus considered in this work.

Results of this classification task can be described using four possible outcomes: Correctly classified phishing messages or True Positives (TP), correctly classified ham messages or True Negative (TN), wrong classified ham messages as phishing or False Positive (FP), and wrong classified phishing messages as ham or False Negative (FN). Given this, machine learning performance criteria, such as Precision and Recall can be constructed. In our work, the F-measure (harmonic mean between Precision and Recall) was used for comparison purposes.

B. Results and Discussions

1) Topic Model Features: In terms of topic model features determined by LDA, the F-measure evaluated over benchmark machine learning algorithms increases as the number of topics gets higher. In this work, up to 25 topics were considered, where 30 words for each topic where used in the Γ feature set (a total of 750 features). In table I, the 10 most relevant words selected for topics 1, 2, 5, 15, 20 and 25 are presented.

TABLE I
Ten most relevant words for five topics extracted by using the LDA topic-model over the phishing corpus.

Topic 1	Topic 2	Topic 5	Topic 15	Topic 20	Topic 25
paypal	account	account	grupo	bank	click
account	messag	fraudul	imagen	account	visa
secur	suspend	bank	cuenta	bankof	card
password	inform	thank	para	america	receiv
protect	termin	suspend	click	wellsfargo	free
inform	warn	fraud	cliente	well	credit
verifi	legal	login	googl	fargo	usernam
click	agreement	secur	bancaria	barclay	success
access	liabil	notif	nuestro	huntington	want
assist	resolv	regard	dato	client	wish

2) Feature Extraction and Selection: By the usage of SVD, it is possible to define the set Υ, which later is combined with Ω and Γ according to 1. In our work, the rank of the VSM matrix was determined as 1780 (less than the total size of vocabulary (25205) and messages (4450)), which represents the total size of of the semantically relevant features for the phishing corpus evaluated. The evaluation of $\Upsilon \cap \Omega$ (recalling that Ω represents the set of relevant keywords), whose cardinality is 405, gives a final set of 377 features. Then, the evaluation of $\Upsilon \cap \Gamma$ (where Γ represents the set of relevant words associated to topics), whose cardinality is 750, gives a final set of 632 features. Finally the evaluation of $(\Upsilon \cap \Gamma) \cup (\Upsilon \cap \Omega)$ gives a final set of 1002 features, which together with Ξ (the set of structural features), the final set is composed by 1017 binary features.

[1] In this work, the set of structural features will be the same as the one presented in [8]

3) Benchmark Algorithms Results: As presented in figure 2, results for all benchmark machine learning algorithms indicates that the feature selection procedure over the $\mathcal{F}$ feature set is the best experimental setup, were all three algorithms achieved their maximum values for the F-measure. It is important to notice that in all single evaluation feature sets (Ω, Υ, Γ and Ξ), the best performance was obtained in the topic-model feature set (Γ). This result was expected, given that the words associated to the topics extracted have a more accurate representation of the malicious set of documents.

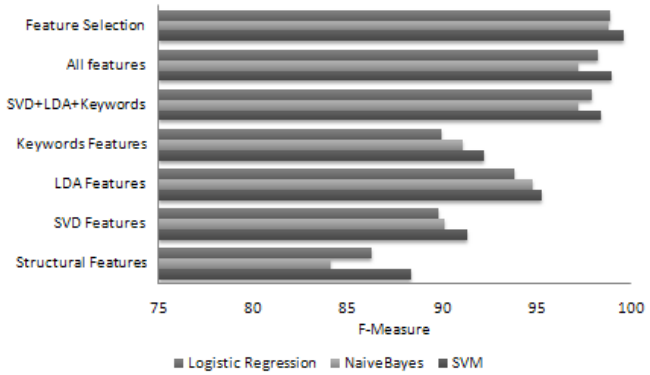

Fig. 2. F-Measure for all benchmark machine learning algorithms evaluated over the different feature sets.

Figure 2 shows that feature selection algorithm is relevant for improving performance measures. F-measure results for the SVM algorithm increases from 99,01% to 99.58%m, for the naïve Bayes increases from a 97.23% to 98.81%, and for logistic regression the F-measures rises from 98.22% to 98.93%. This results are far from their initial evaluation for the simplest feature set (Ξ), where reported F-measures for SVMs, naïve Bayes and logistic regression are 88.24%, 84.12%, and 86.30% respectively.

IV. Conclusion

In terms of characterization of malicious messages, all proposed feature sets reported different levels of performance over the benchmark classification algorithms. The combination of topic-model features and keyword features, filtered by those SVD words and filtered by a mutual information feature selection, presented slightly better results than state of the art feature extraction methods for phishing messages. This characterization is fundamentally based on latent semantic analysis over the email message corpus, where clustering techniques for determining topics and keywords are enhanced by an SVD reduction of the *tf-idf* representation of the corpus. The independent feature sets reported the highest values for the F-measure criteria for the topic-models features (Γ), and the lowest values for the structural features (Ξ).

The machine learning approach for malicious message classification reported interesting results according to the evaluation criteria of benchmark algorithms. The classification performance evaluated over different sets of features, indicates that the SVM algorithm, based on the structural risk minimization, outperforms other machine learning algorithms, such as generative models represented by naïve Bayes, and discriminative classifier models represented by logistic regression. This supports the ussual preference of SVMs for classification tasks, specially in text-mining applications.

Acknowledgment

Support from the Chilean "Instituto Sistemas Complejos de Ingeniera" (ICM: P-05-004-F, CONICYT: FBO16; www.sistemasdeingenieria.cl) and the Chilean Anillo project ACT87 "Quantitative methods in security" (www.ceamos.cl) are greatly acknowledged by the first and third author. The second author gratefully acknowledges the support of the Chilean Computer Emergency Response Team (www.clcert.cl), and CONICYT via Fondecyt 1070332.

References

[1] Saeed Abu-Nimeh, Dario Nappa, Xinlei Wang, and Suku Nair. A comparison of machine learning techniques for phishing detection. In *eCrime '07: Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pages 60–69, New York, NY, USA, 2007. ACM.

[2] Ram Basne, Srinivas Mukkamala, and Andrew H. Sung. *Detection of Phishing Attacks: A Machine Learning Approach*, chapter Studies in Fuzziness and Soft Computing, pages 373–383. Springer Berlin / Heidelberg, 2008.

[3] Andre Bergholz, Jan De Beer, Sebastian Glahn, Marie-Francine Moens, Gerhard Paass, and Siehyun Strobel. New filtering approaches for phishing email. *Journal of Computer Security*, 2009. Accepted for publication.

[4] Andre Bergholz, Jeong-Ho Chang, Gerhard Paass, Frank Reichartz, and Siehyun Strobel. Improved phishing detection using model-based features. In *Fifth Conference on Email and Anti-Spam, CEAS 2008*, 2008.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[7] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 649–656, New York, NY, USA, 2007. ACM.

[8] Gaston L'Huillier, Richard Weber, and Nicolas Figueroa. Online phishing classification using adversarial data mining and signaling games. In *CSI-KDD '09: Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics*, pages 33–42, New York, NY, USA, 2009. ACM.

[9] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: a probabilistic analysis. In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168, New York, NY, USA, 1998. ACM.

[10] Fabrizio Sebastiani. Text categorization. In Alessandro Zanasi, editor, *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, pages 109–129. WIT Press, Southampton, UK, 2005.

[11] Juan D. Velasquez, Sebastian A. Rios, Alejandro Bassi, Hiroshi Yasuda, and Terumasa Aoki. Towards the identification of keywords in the web site text content: A methodological approach. *International Journal of Web Information Systems information*, Vol. 1(1):pp. 53–57, 2005.

[12] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.