A Combined Naïve Bayes and URL Analysis Based Adaptive Technique for Email Classification

Tina R. Patil   Prof. Dr. V. M. Thakare   Prof. Dr. S. S. Sherekar

Abstract **:- Most content based spam filters are rule based or trained off-line. Handling new spam tactics is difficult and prone to high misclassification rate. Incremental adaptive spam mail filtering using Naïve Bayesian classification gives good performance, simplicity and adaptability. Phishing emails contain socially engineered messages to lure victims into performing certain actions, such as clicking on a URL where a phishing website is hosted, or executing a malware code. We model an incremental scheme that receives a stream of emails, and applies the concept of sliding window to train only the last w emails for testing new incoming messages. Subsequently, the new features of tested messages are added to the existing features so that the model will be adaptive to future incoming emails. In this study, we extend the approach to the phishing email classification domain. The primary motive behind this study is that most phishing email messages contain URLs that point to phishing websites, and lexically analyzing the URLs can enhance the classification accuracy of email messages. Our proposed model consists of the combination of the adaptive naïve baysian filter and lexical URL analysis. This model removes oldest emails but keep the features to train new incoming emails. Lexical URL analysis is applied on incoming email after preprocessing. It detects and classify the host website and reports with ham or spam.**

Keywords **: LBS, anonymity, cloaking, location based query processing.**

I. Introduction

Most spam filters analyze the textual part of email messages, and these filters are content-based filters which use text classification. The spam filter algorithms are mostly based on the bag-of-words model. Handling new spam tactics is difficult and prone to high misclassification rate is a social engineering attack that exploits the human using a given system, and therefore user awareness training programs would aid the fight against phishing attacks. Phishing attacks are the social engineering attack through which the victim is persuaded to perform certain actions, such as submitting personal information directly to the phisher, or executing malware.

Keyword extraction can be done to eliminate Phishing fraud detection in cyber crime context [1][2]. To avoid the phishing attacks some text mining methods have been proposed with standard classifier. Model proposed a novel characterization of email using class-topic model. Main contribution of this work is a feature extraction methodology for phishing emails enhancing traditional machine learning algorithms used in email filtering.

II. Background

Most spam filters analyze the textual part of email messages, and these filters are content-based filters which use text classification. The spam filter algorithms are mostly based on the bag-of-words model. The classification of junk emails and Bayesian classification has been widely used [1]. The other well-known approaches are Support Vector Machine (SVM) and k-Nearest. Naive Bayesian method is simple and could be easily implemented as an incremental learning model. Moreover, Naive Bayesian requires linear training time whereas SVM requires quadratic training time and k-Nearest Neighbor requires more testing time [2].

A rule-based spam filters define a set of rules to identify whether an incoming mail is ham or spam. When a new incoming email does not fit with existing rules, the misclassification rate could be high [3]. Spam filters do not evolve as fast as spam techniques. Hence, there is need of spam filter that could be adaptive to meet this challenge. Moreover, incremental spam filtering processes a sequence of incoming emails one at a time in a chronological order, and it differs from batch filtering that uses the whole set of emails [1].

A novel URL tokenization technique was proposed to classify phishing websites. The approach achieved 97% of classification accuracy by lexically analyzing suspect URLs [2]. One of the key objectives of the approach was achieving good classification accuracy while excluding expensive tests that analyze the HTML content of phishing websites [2], or performing networks tests such as DNS queries or Whois lookup over the network.

Different text mining techniques for phishing filtering have been proposed. In research of, 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, proposed a novel characterization of emails using a Class-Topic model. For phishing feature extraction several methodologies have been developed, while for phishing classification data mining approaches have been used [3].

The rest of the paper is organized as: section III and IV describes work done on various methods and existing methodologies. Section V describes detailed analysis of methods. Section VI covers proposed methodologies and section VII describes performance evaluation. Section VIII and IX presents the conclusion and future work of the proposed method.

III. Previous Work Done

In this proposed model, Multinomial Naïve Bayesian with Boolean attributes is used for email classification. Their evaluation combined both rule-based filtering via SpamAssassin and statistical filtering using Bayes filter that have been carried out for the quadratic training time for the SVM and naive baysian classification model. Method has been followed for the incremental baysian algorithm for spam filtering.

Blacklists and heuristics are arguably the most popular phishing detection techniques. Fette et. al. presented the design and evaluation of the first Machine Learning-based email classifier to detect phishing messages, which showed promising results that achieved low false positives while avoiding using blacklists [2][4]. Bergholz et. Al. have proposed a Machine Learning classifier with model-based features — that is, features that themselves are classification models and require to be trained first prior to their use by a parent classifier [5]. The data set used in the evaluation process contained phishing and legitimate samples from publicly available datasets.

Toolan et. Al. proposed a novel Machine Learning ensemble technique that is composed of a parent classifier (C5.0 in this case) with an ensemble of 3 learners (SVM, k-NN with k = 3 and k = 5) that are applied on the legitimate branch of the parent classifier [2][4]. Results are evaluated as a number of feature subset searching methods and features subset evaluators and the outcome was that the Wrapper and Best-first.

Different mining techniques studied by as Logistic Regression, Support Vector Machines (SVMs), and Random Forests are used to estimate classifiers for the correct labeling of email messages while data mining approaches have used for phishing detection. For phishing feature extraction several methodologies have been developed. Bergholz et al. proposed a class-topic model for a novel characterization of email which is used to evaluate the performance of the algorithms [5].

IV. Existing Methodology

Framework has proposed the incremental adaptive spam filtering which can be adapted to new spam pattern. The framework consists of two steps: preprocessing and incremental adaptive spam filtering. In the preprocessing step, a message is tokenized into words. In the filtering step, a filter received those words and classified a message as ham or spam.

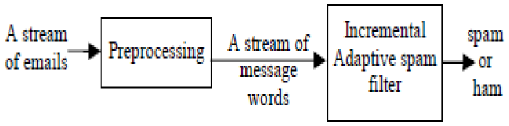

Fig. 1 Framework for spam classification

Incremental Adaptive Spam Filtering: Spam filter introduces the concept of sliding windows, the incremental training and feature updating processes to the model [1]. It consists of two modules as training and testing. Training module perform its work in two steps: feature selection and email classification. In feature selection, the probability ratio of a word *ti* tells the ratio of the probability that the word would be in the spam class to the probability that the word would be in the ham class. While in email classification, The Naïve Bayesian classification based on the Multinomial Naïve Bayesian with Boolean attributes is used for email classification [1].

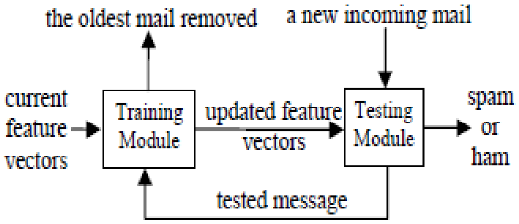

Fig. 2 Incremental Naïve bayes spam filter

Proposed Lexical URL Analysis (LUA) provides a URL tokenization mechanism to export the same logic stated above to the software, so that the software will be able to make better decisions than an average non-technical end-user [2]. LUA is able to discover relations that an educated human would not be able to discover easily, such as: if the keyword signin appears in a specific location the website is legitimate, but if the keyword sign-in appeared in the same location it might be phishing [2].

Lexical URL Analysis (LUA) in provides a URL tokenization mechanism to export the same logic stated above to the software. LUA is able to discover relations that an educated human would not be able to discover easily, [3] such as: if the keyword signing appears in a specific location the website is legitimate, but if the keyword sign-in appeared in the same location it might be phishing. LUA's implementation follows a statistical learning technique to construct a classification model that is able to predict the classes of un-labeled suspect URLs. Phishing probabilities and phishiness is calculated using formula [3].

V. Analysis And Discussion

Results of a baysian framework showed that the number of features has little impact on the ham misclassification rate as it slightly decreases when the number of features is increased for both corpora. On the contrary, the spam misclassification rate slightly increases when the number of features is increased for both corpora. The result corresponds to the fact that larger the window size, more the training emails are in the window, and provides better accuracy. The processing time always increases when the window size increases but not significantly affected by the number of features. Ti is shown that significant improvement of our processing time per message for all corpora.

The processing time per message for Trec06p corpus is about 18 times faster than the batch filter for 1% window size. Results indicate that the number of features has significant effects on spam misclassification rate since it increased when the number of feature is increased. In addition, the spam misclassification rate is higher than the ham misclassification for both varying window sizes and the number of features.

Features set 1-B has an fpr of 1.02%, which is believed to be too high for most production environments. However, the addition of the LUA feature reduced it down to 0.59%.

When RF was run with AdaBoostM1 and using features set 3-A (all of the 48 features including the LUA), its classification model resulted in an f1 score of 99.45%. Only one classifier is known to have a higher f1 score of 99.46%.

Latent semantic gives a focused view on results with tree methods as:

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.

2) Feature Extraction and Selection: By the usage of SVD, it is possible to define the set Y, 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.

3) Benchmark algorithm result: results for all benchmark machine learning algorithms indicates that the feature selection procedure over the F feature set is the best experimental setup, were all three algorithms achieved their maximum values for the F-measure.

VI. Proposed Methodology

The zero-frequency problem could happen when words are present only in ham or spam class. Proposed model maintains only the top selected features. Hence, some old features may be early removed, and would affect the misclassification rate greatly. The addition of URL lexical analysis in phishing email classification is effective and results in a highly accurate antiphishing email classifier. But it does not explore the effectiveness of new features based on time frequency representation of music and lyrics, as well as the hierarchical multilabel classification approach. F-measures are obtained 97% but still do not give the efficient output.

A new proposed method consists of oldest feature extraction in adaptive baysian spam mail filtering with LUA method applied to incoming email.

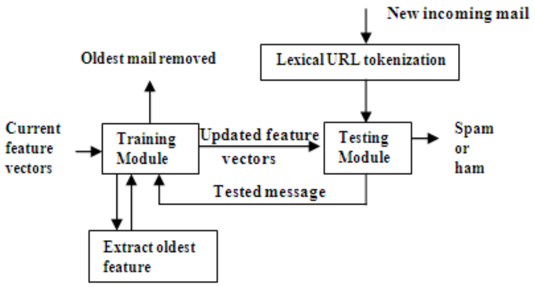

Fig.3 Framework for feature extraction and classification

VII. Possible Outcome And Results

Defining the window size as the percentage of the total size would give a better picture of its effect than expressed it as the actual number of emails in the window size. It would be good to know the optimal window size of each corpus. The processing time per message for Trec06p corpus is about 18 times faster than the batch filter for 1% window size. Results can indicate that the number of features has significant effects on spam misclassification rate since it increased when the number of feature is increased. In addition, the spam misclassification rate is higher than the ham misclassification for both varying window sizes and the number of features.

Ham misclassification rate can be reduced up to 95 %.

VIII. Conclusion

Focus was on the adaptation and the performance issues of an incremental adaptive spam filter. Since the incremental learning is employed in our proposed model, thus the system can be adapted to new spam mails. The sliding window is used to keep the training data up-todate with limited size, and the computed probability ratio is used to reduce the dimension of feature space so that the training time is minimal during the re-training process.

LUA technique enhances the classification accuracy of anti-phishing email filters. LUA feature is primarily focused to classify phishing websites; it may be effective to classify email messages due to the fact that most phishing email messages contain URLs. According to the performance evaluation, the LUA features are effective in enhancing the classifier’s accuracy in all features subsets.

IX. Future Scope

To investigate the issues of optimal window size, develop the dynamic feature selection technique and reduce the number of re-straining process.

For the improved email streaming and decreasing the time, we have to investigate the issues of optimal window size, develop the dynamic feature selection technique and reduce the number of re-training process. Also to extract the features online with mailing application tools is the future work Lexical URL Analysis and to improve the f- measure and recall attributes with latent semantic analysis.

References

[1] Phimphaka Taninpong, Sudsanguan Ngamsuriyaroj,“ Incremental Adaptive Spam Mail Filtering Using Naïve Bayesian Classification”, IEEE 10.1109/SNPD.2009.45, PP 243-248, February 2009.

[2] Mahmoud Khonji, Youssef Iraqi, Andrew Jones,“Lexical URL Analysis for Discriminating Phishing and Legitimate E-Mail Messages,” IEEE 978-1-908320-00, PP 422-427, January 2012.

[3] Gast´on L’Huillier, Alejandro Hevia, Richard Weber, Sebasti´an R´ıos, “Latent Semantic Analysis and Keyword Extraction for Phishing Classification,” IEEE 978-1-4244-6446, PP 129-131, May 2010.

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

[5] 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.