An approach for Malicious Spam Detection In Email with comparison of different classifiers

Umesh Kumar Sah[1] ,Narendra Parmar[2]

[1]M.Tech Scholar, [2] Assistant Professor,
[1,2] Sri Satya Sai College of Engineering, Bhopal

---***---

Abstract— Today, one of the cheapest form of communication in the world is email, and its simplicity makes it vulnerable to many threats. One of the most important threats to email is spam; unsolicited email, especially when advertising agency send a mass mail. Spam email may also include malware as scripts or other executable file. Sometimes they also consist harmful attachments or links to phishing websites. This malicious spam threatens the privacy and security of large amount of sensitive data. Hence, a system that can automatically learn how to classify malicious spam in email is highly desirable. In this paper, we aim to improve detection of malicious spam through feature selection. We propose a model that employs a novel dataset for the process of feature selection, a step for improving classification in later stage. Feature selection is expected to improve training time and accuracy of malicious spam detection. This paper also shows the comparison of various classifier used during the process.

Keywords: email, spam, SVM, Naive Bayes, dataset

I. INTRODUCTION

Today, email is one of the cheapest, fastest and most popular means of communication. It has become a part of everyday life for millions of people for their information sharing [1]. Due to its simplicity email is vulnerable to many threats. One of the most important threats to email is spam: any unsolicited commercial communication. The growth of spam traffic is becoming a worrying problem since it consumes the bandwidth of network, wastes memory and time of users and causes financial loss to both the users and the organizations [2]. Spam also clog up the email system by filling-up the server disk space when sent to many users from the same organization [3].

The most worrying type of spam is malicious spam, which aims to spread numerous emails with links leading to malicious websites. According to Symantec, a sharp rise in malicious URLs at the end of 2014 in comparison to 2013 was related to a change in tactics and a surge in socially engineered spam emails [22]. Unlike cybercrime that targets 'low volume high value' victims such as banks but often requires advanced hacking capability, malicious spam enables malicious content to reach 'high volume low value' targets, which are less likely to have effective anti-virus or other countermeasures in place [4]. In the case of educational institutes, malicious spam threatens the privacy and security of large amount of sensitive data relating to staff and students.

According to [5], certain classes of users, such as executives or military personnel, appear to be targeted together in campaigns of malicious spam. We can hypothesize that the malicious spam emails which are sent to staff in educational institutes share common features. These features have to be explored in order to improve detection of malicious spam in email. In this research, we propose a model that performs feature selection for malicious spam detection in email with the aim of optimizing the classification parameters, the prediction accuracy and computation time for later classification algorithms. A novel dataset will be employed for the process of feature selection, then the set of selected features will be validated using two classifiers: Naïve Bayes (NB), Support Vector Machine (SVM). We believe that our dataset is the first of its kind as no research in the literature was intended to serve malicious spam detection in the field of educational institutes. The terms "spam detection", "spam filtering" and "spam classification" are considered identical and used mutually in this paper.

The rest of this paper is organized as follows: section 2 presents an overview of malicious spam detection. Section 3 summarizes related work and brings out some points in the literature that need more concern. Section 4 illustrates the methodology of our proposed work. Section 5 concludes the paper and points out future work.

II. MALICIOUS SPAM DETECTION: OVERVIEW

A. Malicious Spam Detection

Typically, email systems allow their users to build keyword-based rules that automatically filter spam messages, but most users do not create such rules because they find it difficult to use the software or simply they avoid customizing it [6]. Relying on rule-base filtering is not enough, since the characteristics of spam email (e.g. topics, frequent terms) also change over time as the spammers constantly invent new strategies to beat spam filters. These rules must be constantly tuned by the user which is a time consuming and error-prone process. Hence, a system that can automatically learn how to classify emails is highly desirable [3].

Malicious spam detection is a mature research field with many techniques available such as rule-based, information retrieval based, machine learning based, graph based, and hybrid techniques [7]. Many researchers have used machine learning techniques because email is better classified as spam or ham based on features [8]. Email features are extracted from the email header, subject, body, and the whole email message. The number of features for malicious spam detection can be large, and the inclusion of irrelevant and redundant features can lead to poor classification and high computational overhead. Thus, selecting relevant feature subsets can help reduce the computational cost of feature measurement, speed up learning process and improve model interpretability [8]. Selecting relevant features is called Feature Selection and is presented next.

B. Feature Selection

The number of features present in the feature space (dimensionality) highly affects the efficiency of classification algorithms. The performance of the classifiers tends to degrade as the dimensionality increases [9]. When the dimensionality increases, the required sample size grows exponentially to train a classifier; a problem called “the curse of dimensionality” [10]. Also, high dimensionality takes much training time and classification time [11] [12]. Bishop stated that if complex models (high dimensional models) are trained using data sets of limited size, that will lead to severe overfitting [13]. To overcome these problems, feature selection is suggested to reduce the number of features without loss in classification accuracy. Feature selection will minimize the learning and classification time required by classifier and improve classification accuracy [14].

Feature selection methods consists of two main procedures: subset generation that aims to find the optimal feature subset, and subset evaluation. Subset generation, in turn, is divided into two broad categories: filter-based and wrapper-based. Filter-based methods evaluate features independent of classifier, where wrapper-based methods evaluate feature subsets using the target classifier [15].

C. Importance of The Study

A system that can automatically learn how to detect malicious spam in email is highly desirable for the following reasons:

1) The goal behind malicious spam email is often the acquisition of sensitive information.
2) Since malicious spam emails might evade conventional anti-virus, anti-spam and anti-phishing detection mechanisms, these emails will exhaust the limited resources in our system like storage space and network bandwidth.
3) Usually, malicious spam emails are created such that they are relevant to the recipient. Email addresses, subject lines and content are tailored to increase the interest of the intended target attracting them to open the email, this will lead to waste valuable time of the user.

The proposed study aims at improving malicious spam detection methods for email systems. The issue of malicious spam detection in email system will be investigated.

III. RELATED WORK

A. Literature Review

Many researchers employed machine learning algorithms for spam detection. Other techniques such as blacklisting and heuristic techniques are excluded from this section as they have high false positive rates. Authors in [16] used Bag of Words (BoW) features and SVM as a feature selection strategy. a MapReduce-based parallel SVM algorithm was presented in [17] for fast spam filter training. To mitigate accuracy degradation in classification, the parallel SVM was augmented with ontology semantics. Artificial Neural Networks (ANN) were also used for spam classification in [18] due to their flexible structure and non-linearity transformation to accommodate latest spam patterns. In their study, [19] aimed to improve the weight assignment in ANN for spam detection using Genetic Algorithm (GA). Authors in [20] proposed QUANT (QUAdratic neuron-based Neural Tree), which is a tree-structured neural network composed of quadratic neurons at the decision nodes of the tree. In addition, it provided a structured approach to decide the architecture of a neural network.

Another widely used machine learning algorithm for spam detection is Random Forest (Random Trees). Authors in [21] proposed an optimal spam detection model based on Random Forests (RF) to enable parameters optimization and feature selection. They provided the variable importance of each feature to eliminate the irrelevant features. Furthermore, they decided an optimal number of selected features during overall feature selection, and also in every feature elimination phase. In [7], rich descriptive sets of text features were presented for the task of identifying emails with malicious attachments and Uniform Resource Locators (URLs). Random Forest was applied for feature selection. Authors in [3] developed classification methods using persistent threat and recipient oriented features, designed to detect Targeted Malicious Email (TME). Random Forest was employed to select the top 10 features which best separated TME and non TME. Two alternative forms of random projections were proposed and compared in [8]: The Random Project method that employed a random projection matrix to produce linear combinations of input features, and the Random Boost method which is a combination of Logit Boost and Random Forest. Random feature selection was employed to enhance the performance of the Logit Boost algorithm.

Other popular learning algorithms that have been applied to spam detection include Naïve Bayes [23], AdaBoost [11], Bayes Net (BN) [11], Negative Selection Algorithm (NSA) [13], Symbiotic Filtering [17], Fuzzy Classification [20] and Rough Sets [19].

B. Remarks on Literature Review

Based on the previous review, several remarks can be pointed out as follows:

1) Most researches on malicious spam detection tended to implement content-based approaches, where content-based features were used to train machine learners to detect spam. Number of features varied from 38 to 83 features divided into 4 main categories: Header features, Subject features, Payload (body) features and Attachment features.
2) The most frequent machine learning techniques used in malicious spam detection were SVM, ANN, RF, Naïve Bayes and AdaBoost.
3) A number of researches presented hybrid approaches combining known machine learning techniques with algorithms
inspired by Artificial Immune Systems like Negative Selection, or employed rule-based approaches like Rough Sets.
4) Datasets (emails) used in the experiments in the literature were public, generic and relatively old like Spambase (1999), SpamAssassin (2006) and TREC (2007). Among the researches in this review, only one research on 2015 carried out experiments on dataset from 2013.
5) Almost no research was intended to serve malicious spam detection in a specific domain or field. Most researches proposed generic systems. Our research will focus on the educational field specifically employing a novel dataset to investigate whether or not features for malicious spam detection that were presented in the literature could be reduced for similar systems in the educational field. We will also explore new features (if exist) for malicious spam detection in the educational field that were not presented in the literature.

IV. PROPOSED METHODOLOGY

Text mining (deriving information from text) is a wide field which has gained popularity with the huge text data being generated. Automation of a number of applications like sentiment analysis, document classification, topic classification, text summarization, machine translation, etc has been done using machine learning models.

Spam filtering is a beginner's example of document classification task which involves classifying an email as spam or non-spam (a.k.a. ham) mail. Spam box in your Gmail account is the best example of this. So let's get started in building a spam filter on a publicly available mail corpus. I have extracted equal number of spam and non-spam emails from Ling-spam corpus.

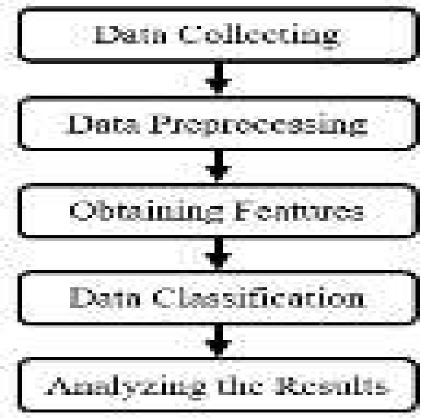

Figure 1. Proposed Architecture for Malicious Spam Detection.

We will walk through the following steps to build this application :

1.Preparing the text data.
2.Creating word dictionary.
3.Feature extraction process
4.Training the classifier

Further, we will check the results on test set of the subset created.

1. Preparing the text data.

The data-set used here, is split into a training set and a test set containing 702 mails and 260 mails respectively, divided equally between spam and ham mails. You will easily recognize spam mails as it contains *spmsg* in its filename.

In any text mining problem, text cleaning is the first step where we remove those words from the document which may not contribute to the information we want to extract. Emails may contain a lot of undesirable characters like punctuation marks, stop words, digits, etc which may not be helpful in detecting the spam email. The emails in Ling-spam corpus have been already pre-processed in the following ways:

a) Removal of stop words – Stop words like "and", "the", "of", etc. are very common in all English sentences and are not very meaningful in deciding spam or legitimate status, so these words have been removed from the emails.
b) Lemmatization – It is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. For example, "include", "includes," and "included" would all be represented as "include". The context of the sentence is also preserved in lemmatization as opposed to stemming (another buzz word in text mining which does not consider meaning of the sentence).

We still need to remove the non-words like punctuation marks or special characters from the mail documents. There are several ways to do it. Here, we will remove such words after creating a dictionary, which is a very convenient method to do so since when you have a dictionary, you need to remove every such word only once.

2. Creating word dictionary.

It can be seen that the first line of the mail is subject and the 3rd line contains the body of the email. We will only perform text analytics on the content to detect the spam mails. As a first step, we need to create a dictionary of words and their frequency. For this task, training set of 700 mails is utilized. This python function creates the dictionary for you.

Dictionary can be seen by the command print dictionary. You may find some absurd word counts to be high but don't worry, it's just a dictionary and you always have the scope of improving it later. If you are following this blog with provided data-set, make sure your dictionary has some of the entries given below as most frequent words. Here I have chosen 3000 most frequently used words in the dictionary.

3. Feature extraction process.

Once the dictionary is ready, we can extract word count vector (our feature here) of 3000 dimensions for each email of training set. Each word count vector contains the frequency of 3000 words in the training file. Of course you might have guessed by now that most of them will be zero. Let us take an example. Suppose we have 500 words in our dictionary. Each word count vector contains the frequency of 500 dictionary words in the training file. Suppose text in training file was "Get the work done, work done" then it will be encoded as [0,0,0,0,0,…….0,0,2,0,0,0,……,0,0,1,0,0,…0,0,1,0,0,……2,0,0, 0,0,0]. Here, all the word counts are placed at 296th, 359th, 415th, 495th index of 500 length word count vector and the rest are zero.

The below python code will generate a feature vector matrix whose rows denote 700 files of training set and columns denote 3000 words of dictionary. The value at index 'ij' will be the number of occurrences of jth word of dictionary in ith file.

4. Training the classifiers

Here, I will be using scikit-learn ML library for training classifiers. It is an open source python ML library which comes bundled in 3rd party distribution anaconda or can be used by separate installation following this. Once installed, we only need to import it in our program.

I have trained two models here namely Naive Bayes classifier and Support Vector Machines (SVM). Naive Bayes classifier is a conventional and very popular method for document classification problem. It is a supervised probabilistic classifier based on Bayes theorem assuming independence between every pair of features. SVMs are supervised binary classifiers which are very effective when you have higher number of features. The goal of SVM is to separate some subset of training data from rest called the support vectors (boundary of separating hyper-plane). The decision function of SVM model that predicts the class of the test data is based on support vectors and makes use of a kernel trick.

Checking Performance

Test-set contains 130 spam emails and 130 non-spam emails. If you have come so far, you will find below results. I have shown the confusion matrix of the test-set for both the models. The diagonal elements represents the correctly identified(a.k.a. true identification) mails where as non-diagonal elements represents wrong classification (false identification) of mails.

Multinomial NB	Ham	Spam
Ham	129	1
Spam	9	121
SVM(Linear)	Ham	Spam
Ham	126	4
Spam	6	124

Both the models had similar performance on the test-set except that the SVM has slightly balanced false identifications. I must remind you that the test data was neither used in creating dictionary nor in the training set.

V. CONCLUSION AND FUTURE WORK

This work proposes a model for improving detection of malicious spam in email. Our model will employ a novel dataset for the process of feature selection, then validate the set of selected features using three classifiers known in spam detection: Naïve Bayes, Support Vector Machine and Multilayer Perceptron. Feature selection is expected to improve training time and accuracy for the classifiers.

REFERENCES

[1] S. Whittaker, V. Bellotti and P. Moody, “Introduction to this special issue on revisiting and reinventing e-mail”, Human-Computer Interaction, 20(1), 1-9, 2005.

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

[3] I. Koprinska, J. Poon, J. Clark and J. Chan, “Learning to classify e-mail”, Information Sciences, 177(10), 2167-2187, 2007.

[4] M. Alazab and R. Broadhurst, “Spam and Criminal Activity”, Australian Institute of Criminology, 2014.

[5] R. Amin, J. Ryan, and J. van Dorp, “Detecting Targeted Malicious Email Using Persistent Threat and Recipient Oriented Features”, IEEE Secur, Priv. Mag, (99), 1-1, 2012.

[6] X. Carreras and L. Màrquez, “Boosting trees for anti-spam email filtering”, Proceedings of 4th International Conference on Recent Advances in Natural Language Processing, 2001.

[7] K. N. Tran, M. Alazab and R. Broadhurst, “Towards a Feature Rich Model for Predicting Spam Emails Containing Malicious Attachments and URLs”, 11th Australasian Data Mining Conference, Canberra, 2015.

[8] F. Temitayo, O. Stephen and A. Abimbola, “Hybrid GA-SVM for efficient feature selection in e-mail classification”, Computer Engineering and Intelligent Systems, 3(3), 17-28, 2012.

[9] J. Thomas, N. S. Raj and P. Vinod, “Towards filtering spam mails using dimensionality reduction methods”, Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference- (pp. 163-168), IEEE, 2014.

[10] M. Verleysen and D. François, The curse of dimensionality in data mining and time series prediction, In Proceedings of the 8th International Work-Conference on Artificial Neural Networks (IWANN), Barcelona, Spain, 2005, 758–770

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, “The WEKA data mining software: an update”, ACM SIGKDD explorations newsletter, 11(1), 10-18, 2009.

[12] L. C. Molina, L. Belanche and À. Nebot, “Feature selection algorithms: a survey and experimental evaluation”, Data Mining, 2002, ICDM 2003 Proceedings, 2002 IEEE International Conference on (pp. 306-313), IEEE, 2002.

[13] R. Amin, J. Ryan, and J. van Dorp, “Detecting Targeted Malicious Email Using Persistent Threat and Recipient Oriented Features”, IEEE Secur, Priv. Mag, (99), 1-1, 2012.

[14] Y. S. Hwang, “Wrapper-based Feature Selection Using Support Vector Machine”, Life Science Journal, 11(7), 2014.

[15] H. Wang, D. Bell and F. Murtagh, “Axiomatic approach to feature subset selection based on relevance”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 21(3), 271-277, 1999.

[16] S. Ali, S. Ozawa , J. Nakazato, T. Ban, and J. Shimamura, “An Online Malicious Spam Email Detection System Using Resource Allocating Network with Locality Sensitive Hashing”, Journal of Intelligent Learning Systems and Applications, 7, 42-57, 2015.

[17] G. Caruana, M. Li, and Y. Liu, “An ontology enhanced parallel SVM for scalable spam filter training”, Neurocomputing, 108, 45-57, 2013.

[18] K. L. Goh, A. K. Singh and K. H. Lim, “Multilayer perceptrons neural network based web spam detection application”, Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on (pp. 636-640), 2013.

[19] A. Arram, H. Mousa and A. Zainal, “Spam detection using hybrid Artificial Neural Network and Genetic algorithm”, Intelligent Systems Design and Applications (ISDA), 13th International Conference on (pp. 336-340), IEEE, 2013.

[20] M. C. Su, H. H. Lo and F. H. Hsu, “A neural tree and its application to spam e-mail detection”, Expert Systems with Applications, 37(12), 7976-7985, 2010.

[21] S. M. Lee, D. S. Kim, J. H. Kim and J. S. Park, “Spam detection using feature selection and parameters optimization”, Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International.

[22] Symantec, Internet Security Threat Report 2015, 2015.

[23] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras, “Spam Filtering with Naive Bayes – Which Naive Bayes?”, CEAS 2006 - Third Conference on Email and Anti-Spam, July 27-28, 2006, Mountain View, California USA.