Re – Tokenization of spam Mails

Chandra Srikanth Prathi[1], Saveetha D[2]

[#]Department of IT-Information Security and Computer Forensics, SRM University
Kattankulathur, Chennai, TN., India
[1]chandraprathi@gmail.com

***Abstract*— a Spam mail has become an important issue in the field of e-mailing system. More than 80% of the mails that receiving today around the world are spam mails. The existing spam filters like Bayesian spam filter has failed with the implementation of tokenizing the mails before sending them to the victim. In these the attackers are inserting some special symbols in-between the text in order to make the body of the mail to be unreadable by the spam filters. This paper will details with how to de-tokenize those attacks in the mails and make it to the original format without missing the information.**

***Keywords*— unreadable format, Bayesian spam filter, Vector Space Model, De-tokenization**

I. INTRODUCTION

Due to the globalization electronic mailing system has been increased to the extreme level. Public mailing systems like Gmail, yahoo, rediff etc., has implemented the security firms to provide a secure mailing system but still spam mails are creating an enormous issues in the field of information security. Because of these issues the sensitive information is leaking from each and every individuals to the organization level. To overcome these issues engineers from all the world have implemented the algorithms named as spam filters. One of the famous spam filters like Bayesian spam filter is using now-a-days in filtering the spam mails that are sending by the attackers to their victims to implements attacks like phishing, attaching virus codes, fooling them by fake jackpots etc. The Bayesian spam filter algorithm is filtering the spam mails by using the vector space model and providing the only legitimate mails to the end users. For this reason all the public mailing systems are using this spam filter in their organizations. Not only in the public mailing organizations, but for the intranet purpose to the private organizations are implementing the same in their mailing configuration.

II. APPROACH

Whenever we send a mail in the public mailing system based on the vendor, the mail may encrypted or not. Every that we send to the recipient will reach the central server of that organization and be saved in the recipient database in the same server (if the recipient is also a customer to that organization) or may be transferred to the different server (if the recipient belongs to other vendor).

But every time, the mail after reaching its central server will undergoes the classification of legitimate mail or not by the vendor's customized spam filter system. There it will be classified as a 'spam' or 'ham' mail based on the content / format of the mail. The methods which are used to send the spam mails through the mailing system has compromised, with the spam filters. But the attackers are improving their tactics in order to implementing the spam mails, they are using the tokens in-between the texts making them unreadable by the spam filters. By which they are compromising the present spam filters.

The objective of this paper is to retokenize the tokenized texts in the body of the mail by implementing an algorithm called 'D-Tokenization'. This can be done by parsing the string or the text in the body of the mail to find the ‘ngrams’ and matching with the dictionary of words to make it into the original format. In addition to this if we implement this approach before sending the mail itself, we can save a huge bandwidth that is wasted due to the spam mails daily throughout the world.

III. BAYESIAN SPAM FILTER

Bayesian spam filter will use the probability in calculating the mail as a spam or ham. In classifying the mail into the two categories (‘spam’ or ‘ham’), the algorithm discards the info like recipient’s mails id and some unwanted words like are, the, a, an etc. because with these words into the consideration, there is not that much influence of classifying the mails instead of increasing the time of complexity of the algorithm. So after removing the unwanted words, we will prepare a vector of words and calculate the occurrence of words in the vector and accordingly calculate the probability of the each and every word in ‘spam’ and ‘ham’ respectively. Here Pr(W|S) and Pr(W|H) gives the probability of values both in spam and ham respectively. And coming to Pr(S) and Pr(H) gives overall ‘spam’ and ‘ham’ probability of the mail. With the consideration of all these values we calculate the ‘spam’ and ‘ham’ probability of the body of the mail and if the value of the ‘spam’ is higher than ‘ham’ then the mails will be declared as ‘spam’ mail. Below is the formula in categorizing the mail as a ‘spam’ or ‘ham’

$$\Pr(S|W) = \frac{\Pr(W|S)\cdot\Pr(S)}{\Pr(W|S)\cdot\Pr(S)+\Pr(W|H)\cdot\Pr(H)}$$

IV. Vector space model

Vector space model is used to represent the string of words in vector, an algebra model as vector of documents. This model is used to give the ranking to the documents according to their occurrence in a document. In this model the documents is divided into no. of documents and queries will be applied on documents for further process. The documents and queries will applied as follows:

$$d_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$
$$q = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$$

Advantages:

1. Simple model based on linear algebra
2. Terms weights not binary
3. Allow partial marking

Disadvantages:

1. Weighting is intuitive but not very formal
2. Theoretical assumes terms are statistically independent.

V. Unreadable format

Making the string as an unreadable format, by inserting some special symbols in-between characters makes the string the unreadable format for the system or algorithm to recognize. So when the algorithm apply on those string formats, algorithm will fails to recognize the strings and pass as a ‘ham’. So in this way the attackers are compromising the present spam filter algorithms.

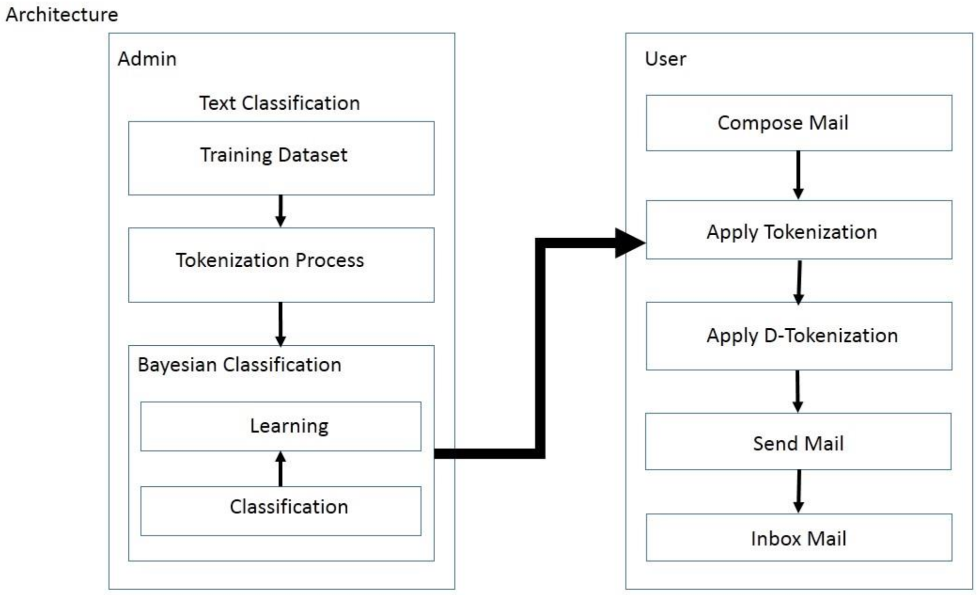
The process of changing the string into unreadable format is as follows:

1. Cut the string into individual strings based ‘split’ method.
2. Select the word size of greater than 3 count and select the substring of 0th position to random position less than the word size.
3. Insert the special symbol and again combine the string with the remaining part. If the word size is little bit high for ex: extraordinary, single insertion of special symbol don’t makes the string to unreadable format, so we need to make multiple substrings and insert 2 or more special symbols into it.

VI. Architecture

Above is the architecture of the proposed system, here initially the system should be trained to classify the ‘spam’ and ‘ham’. For this thing initially the admin or organization will train the system by classifying them into ‘ham’ and ‘spam’. The more no. of datasets that we are training the more efficient the algorithm will works.

Here, the unreadable format method is used for the demo of the algorithm which will make the strings to their original format. After applying the algorithm proposed in this paper, the string of the unreadable format will changed to their original format. Below is the process of how it works,



1. Initially the system is trained to make the individual words from the strings
2. Secondly recognize the words for unreadable format, here there should be a method of recognizing whether the string is in readable or unreadable format.
3. This will increase the performance of the algorithm, and finally splitting according to the special symbol.
4. After splitting the words, we should apply the ngrams on that, and check for the relevant words in the dictionary of words (which we should prepare in advance). If a words matches with a words, that's it D-Tokenization has been completed.
5. After completion of all the above process we collect all the readable format strings again and make them into a single string which is our body of the mail.
6. In this way we can compromise the attacks that are implementing by the attackers.

VII. CONCLUSION AND FUTURE ENHANCEMENT

Finally we can conclude that spam is notorious issue that is facing by the present world, why because with the help of spam mails, the attackers can implement various types of attacks like phishing, social engineering, malwares etc., but with this new type of unreadable type of attacks, the world's famous spam filters like Bayesian spam filter itself has been compromised. So with this new type D-tokenized method applying before Bayesian spam filter, this attacks can be counter measured.

The future enhancement of this algorithm will be using some localization of words (names) in the dictionary like Srikanth, Venkatesh, and Someswara Prasad, will gives the algorithm to re-format each and every string to their original format.

REFERENCE

[1] Machine Learning in Automated Text Categorization, Fabrizio Sebastiani Consiglio Nazionale delle Ricerche, Italy.

[2] Naive Bayes Spam Filtering Using Word-Position-Based Attributes, Johan Hovold Department of Computer Science Lund University Box 118, 221 00 Lund, Sweden.

[3] Designing and Conducting Phishing Experiments Peter Finn Markus Jakobsson Dept. of Psychology School of Informatics Indiana University Bloomington, IN 47406

[4] JURD: Joiner of Un-Readable Documents algorithm for Reversing the Effects of Tokenisation Attacks against Content-based Spam Filters, Igor Santos, Carlos Laorden, Borja Sanz, Pablo G. Bringas S3Lab, DeustoTech – Computing Deusto Institute of Technology, University of Deusto Avenida de las Universidades 24, 48007, Bilbao, Spain.

[5] Learning to Filter Spam E-Mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach, Ion Androutsopoulos., Georgios Paliouras., Vangelis Karkaletsis., Georgios Sakkis., Constantine D. Spyropoulos. and Panagiotis Stamatopoulos. .Software and Knowledge Engineering Laboratory Institute of Informatics and Telecommunications National Centre for Scientific Research gDemokritosh 153 10 Ag. Paraskevi, Athens, Greece.