ISSN No. 0976-5697

Volume 4, No. 8, May-June 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An improved Bayesian filtering technique for spam recognition

Ainam Jean-Paul
Department of Computer Science
Babcock University
Ilishan-Remo, Ogun State, Nigeria
jpainam@gmail.com

Adekunle Y.A
Department of Computer Science
Babcock University
Ilishan-Remo, Ogun State, Nigeria
adekunleya@gmail.com

Adio A.K
Department of Basic Sciences
Babcock University
Ilishan-Remo, Ogun Stage, Nigeria
Adesinaadio1@gmail.com

Abstract: In this paper, we presented models and software for spam recognition using an improved Bayesian filtering technique. Based on a corpus from Androutsopoulos et al, our Spam Recognition framework outperforms other state-of-the-art learning methods based on Bayesian algorithm in terms of spam detection capability. Our software has proved an accuracy of 99.9% of good classification. The 0.1% of other messages have been classify as "may be spam" due to their vagueness signature. Brief, in the case of extremely high misclassification cost, our model still remains stable accuracy with low computation cost, while other methods' performance deteriorates significantly as the cost factor increases.

Keywords: Spam filter; filter technique; Bayesian algorithm; Bayes technique; Naïve Bayes; Spamming techniques; Spamming preventive techniques.

I. INTRODUCTION

Spam Recognition is field in the Pattern Recognition domain aiming to develop new functions in order to identify the category allotted to incoming e-mail, which can be spam or legitimate. Spamming is the abuse of electronic messaging systems to send unsolicited bulk messages. In this paper, we focus on Spam Recognition using the Bayesian algorithm.

Bayesian algorithm is a statistical algorithm based on content-based and learning. Spam emails normally relate to specific topics such as prescription drugs, get-rich-quick schemes, financial services, qualifications, online gambling, discounted or pirated software [1]. With a huge volume of spam messages received every day, it would not be practical for human users to detect spam by reading all of them manually. Therefore there is a need of building software which will perform this task without the help of human being and which can improve the accuracy with time: that is Computer Learning.

Bayesian spam filtering is a very powerful technique for dealing with spam, that can tailor itself to the email needs of individual users, and gives low false positive spam detection rates that are generally acceptable to users. However, it can present inaccuracy in classification. Thus, there is a need of improving this technique. Several works have been done in this angle, but all remains inaccurate.

The aim of this investigation is to analyse Bayesian algorithm applied in Spam Recognition. We start by providing a model of the application, from the conception to the design. The software is able to recognize spam email without any human intervention after an effective set of training. We conclude by underlying possible further studies for this area.

II. RELATED WORKS

Here are some of the related works in Bayesian anti-spam:

Vikas P. et al, An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques. [2]

They examined`` the effectiveness of statistically-based approaches Naïve Bayesian anti-spam filters and designed a derivative filter based on relative numbers of tokens. They train the filter using a large corpus of legitimate messages and spam and test the filter using new incoming personal messages. Finally, they look at the effectiveness of the technique, and evaluated different threshold values in order to find an optimal anti-spam filter configuration without specifying how software can use the new configurations. They just conclude that additional safety precautions are needed for Bayesian anti-spam filter to be put into practice.

Trevor Stone: Parameterization of Naïve Bayes for Spam Filtering, University of Colorado 2003. [3]

He presents the results of applying the Naïve Bayes algorithm to the problem of filtering unwanted junk or "spam" email. He then, explores the effect of several parameters including corpus size and feature extraction methods. Finally, he compares his results to several published statistical spam-filtering approaches.

Vangelis Metsis et al, Spam Filtering with Naïve Bayes – Which Naïve Bayes? [4]

Here, they noted that there are several forms of Naïve Bayes and declared that literatures do not always acknowledge that fact. They discussed five different versions of Naïve Bayes and compared them on six new, non-encoded datasets that contain ham messages of particular Enron users and fresh spam messages. For that, they adopted an experimental procedure that emulates the incremental training of personalized spam filters and plotted curves that allowed them to compare the different versions of Naïve Bayes over the entire tradeoff between true positives and true negatives.

Ion Androutsopoulos et al, "An Evaluation of Naïve Bayesian Anti-Spam Filtering". [5]

They conduct a thorough evaluation using a corpus that they made publicly available. At the same time, they investigate the effect of attribute-set size, training-corpus size, lemmatization and stop-lists on the filter-s performance, issues that had not been previously done. They introduce appropriate cost-sensitive evaluation measures and reach the conclusion that additional safety nets are needed for the Naïve Bayesian anti-spam filter to be viable in practice.

III. TECHNIQUE OF SPAMMING

Spamming Techniques refer to whole techniques used by an offensive user on Internet to send unsolicited E-mail messages to people in order to harm them. These techniques are becoming more and more sophisticated due to the increasing popularity and low cost of E-mail. Spam is usually classified into two categories which have different effects on Internet users:

- Cancellable Usenet Spam which is a single message sent at many Usenet newsgroups. This spamming attack can overwhelm the users with a barrage of advertising or other irrelevant posts; and
- Email Spam which targets individual users with direct mail messages.

Though there are different types of spam, they all share some common proprieties such as:

- First, sender's identity and address are concealed,
- Second, spams are sent to a large number of recipients and in high quantities, and
- Finally spams are unsolicited.

These different types of Spam message are successfully achieved if the Spammers underwent two mains steps: collect E-mail address and bypass anti-pass measures.

A. Collect E-mail Address

The techniques used by the spammers to collect these E-mail include: [1]

- Offers, discount;
- Blog or forum comments;
- Buy a list of addresses from web sites;
- Steal users address books on compromised computers;
- Guess email addresses and then send email to see if it goes through;

All those techniques consist of using false reasons to trick a user into giving up their email address.

B. Bypass Anti-spam Measures

So, after getting valid E-mail addresses of potential target victims by using one or another of the aforementioned techniques, the spammers should cleverly bypass anti-spam measures in order to get finally the victims. This task involves disguising the Spam as a non-spam message with normal appearing subject lines and other ways of getting around anti-spam software. The next section describes some techniques used to prevent Spam

IV. PREVENTIVE TECHNIQUE

Various countermeasures to Spam have been proposed to mitigate the impacts of unsolicited emails. These techniques include [6]:

- Hiding contact information;
- Looking at filtering software;
- Basic structured text filters;
- Whitelist or verification filters;
- Bayesian word distributed filters.

Some of these aforementioned techniques make use of Spam recognition methods to prevent E-mail spams. Some of them that make used of straightforward techniques such as whitelist and text filters are very ineffective because a user can forget to add a well-known email address to his whitelist or assuming that an approved user changes his email address. In one or other way, the legitimate message becomes spam. Also, text filters are very limited due to the changing nature of spam messages and to their text-based contents. In fact, text filters can fail because the actual content of an unsolicited message does not always makes a message becomes a spam. For these points of view, spamming becomes a very challenging problem to the sustainability of the Internet, given the content of emails the only foundation for spam recognition.

V. SPAM RECOGNITION METHODS

Pattern Recognition field has evolved due to recent advances in Machine Learning Techniques. These advances have attracted immense attention from researchers to explore the applicability of learning algorithms in anti-spam filtering. Yet, Spam Recognition is a procedure that can, without human intervention, classify a new incoming email as spam or legitimate according to the knowledge collected from the training stage. Many algorithms or methods for spam recognition have been developed to achieve spam filtering. This section discusses commonly used learning algorithms for spam recognition problems (six algorithms). [1]

- Memory Based Learning simply consists of storing the training messages and then classified the incoming messages as spam or non-spam by estimating their similarity to the stored examples.
- Boosted Decision Tree is a well-known technique in machine learning; it uses a decision tree as a predictive model which maps observations from the instance space to the target values.
- Artificial Neural Network is a collection of interconnected nodes or neurons. The best known example of one is the human brain, the most complex and sophisticated neural network.

The next section is specific to our study. It provides a thorough examination for Bayes spam technique and proposed an improved technique based on it.

VI. BAYESIAN TECHNIQUE

Bayesian Spam filtering is a technique of e-mail filtering that makes use of Naïve Bayes Classifier to identify Spam E-mail. Bayes Classifier was the first method used in the mail-filtering program released in 1996 by Jason Rennie's file. Bayesian classifiers work by correlating the use of words with spam and non-spam emails and then using Bayesian inference to calculate a probability that an email is or is not spam. In fact, Particular words have particular probabilities of occurring in spam email and in legitimate email. [6].

A. *Procedure in bayesian filtering technique*

1) Learning set and priors

The first step in Bayesian filter technique is learning. That is, the filter must first be trained so it can build up in advance the probability that a given word appears in spam email. That means, for all words in each training email, the filter will adjust the probability that each word will appear in spam or legitimate email in its database.

2) Compute Probability or features

After training, Bayesian filters used Bayes' Theorem to compute the probability that an email with particular set of words in it belongs to either category. Each word in the email contributes to the email's spam probability, or only the most interesting words which consist of removing any other words that suggest a list such as 'and, or, whether, either …'. Bayes Theorem is defined as followed:

$$\Pr(S|W) = \frac{\Pr(W|S).\Pr(S)}{\Pr(W|S).\Pr(S) + \Pr(W|H).\Pr(H)} \quad [7] \qquad (1)$$

Where the symbols in equation (1) are defined as fellow:

- Pr(S|W) is the probability that a message is a spam, knowing that the word W is in it.
- Pr(S) is the overall probability that any given message is spam;
- Pr(W|S) is the probability that the word W appears in Spam messages;
- Pr(H) is the overall probability that any given message is not Spam (is "Ham);
- Pr(W|H) Is the probability that the word W appears in Ham messages.

The Bayes formula seem to be long to used, thus most Bayesian Spam detection software makes the assumption that there is no prior reason for any incoming message to be Spam rather than Ham, and considers both cases to have equal probabilities of 50% i.e.$\{\Pr(S) = 0.5;\ \Pr(H) = 0.5\}$. Therefore, this assumption permits simplifying the general formula to:

$$\Pr(S|W) = \frac{\Pr(W|S).\Pr(S)}{\Pr(W|S)+\Pr(W|H)} \quad [7] \qquad (2)$$

This quantity is called "spamicity" or "spaminess" of the word W and can be computed. The number Pr(S|W) used in equation (2) is approximated to the frequency of messages containing in the messages identified as spam during the learning phase.

3) Decision or validation

After computing the probability of each word, the Bayesian filter computes probability that a message is spam by taking into consideration all of its words (or a relevant subset of them), if the total exceeds a certain threshold (say 95% or 80%), the filter will mark the email as a spam.

The following diagram presents the steps followed by the Bayes Technique to classify the incoming messages.

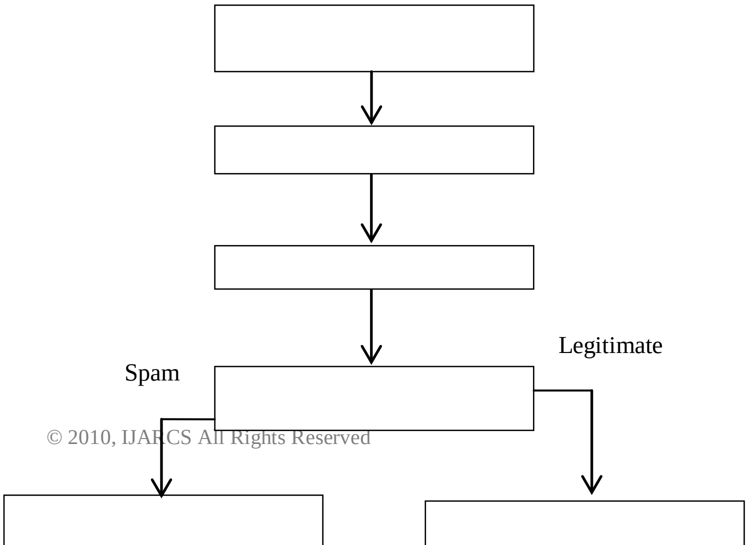

Figure 1. Model of naîve Bayesian Spam filtering

B. *Bayesian shortcomings*

Bayesian Spam filtering technique make the naïve assumption that the words present in the message are independent event. That is wrong in natural languages like French, where the probability of finding an adjective, for example, is affected by the probability of having a noun. Such assumptions make the spam filtering software a naïve Bayes classifier which can be improved by combining individual probabilities to Bayes Theorem. Another disadvantage of this technique is that it is unable to analyse picture which would contain sensitive words classified as spam. However Google has proposed a more efficient solution by performing an OCR (Optical Character Recognition) [8] which analyse the text inside. This technique is used by its Gmail email system.

Bayes filter can be considered as inefficient in the case that, if a word has never been met, both denominator and numerator are equal to zero in Bayes formula and the filter can decide to discard such words for which there is no information available, which is somehow wrong. In addition, one of the disadvantages of Bayes Technique model is that the number of "words" in email is virtually unbounded.

VII. OUR PROPOSED MODEL

The Bayesian Filtering Technique presents a lot of shortcoming which can be improved by using a Data Structure called Graph.

A. *Data structure: Graph*

In computer Science, a Graph is an Abstract Data Type that is meant to implement the graph and hyper graph concepts from mathematics. A Graph data Structure consists of a finite (and possibly mutable) set of ordered pairs, called edges or arcs, of certain entities called nodes or vertices. As in mathematics, an edge (x, y) is said to point or go from x to y. the nodes may be part of the Graph Structure, or may be external entities represented by integer indices or references. A Graph Data Structure may also associate to each edge some edge value, such as a symbolic or a numeric attribute (cost, capacity, length, etc.)

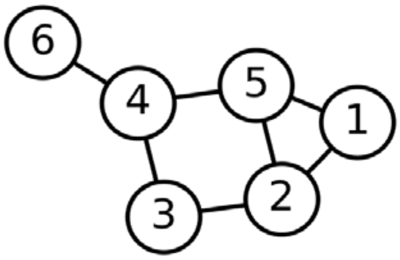

Figure 2. A data structure Graph. [8]

In a Graph, when the edges have a direction, the graph is called a directed graph or digraph, and the edges are called directed edges or arcs. Here, we shall be exclusively concerned with directed graphs, and so when we refer to an edge, we mean a directed edge. This is not a limitation, since an undirected graph can easily be implemented as a directed graph by adding edges between connected vertices in both directions. The following diagram shows a graph with 5 vertices and 7 edges. The edges between A and D and B and C are pairs that make a bidirectional connection, represented here by a double headed arrow.

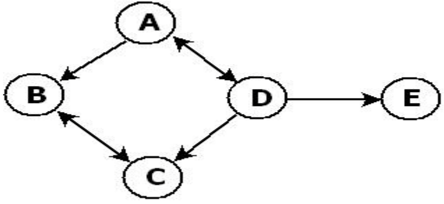

Figure 3. An oriented Graph. [8]

Within the context of this study, we are going to use an oriented graph to represent our data.

The Bayesian Filtering Technique has shown many shortcomings as respect to spam classification. This disadvantage is due to the fact that the Bayesian Filtering make the assumption that the words present in the message are independent event. That means, Bayes computes in advance the probability that a given word appears in spam email or not. To improve this technique, we propose a new way to compute the probability of each word. The steps follow to perform this task include:

a) Construct a whole Dictionary of words present in the message: Instead of determining the probability of each word separately, we consider the whole sentence where the word appears. A sentence in common language like English and French, starts in one point and always ends by a full stop. That is the easily way to recognize a sentence. Base on this, the Technique starts by building a dictionary of possible word that may come along with each word. It is where the Data Structure Graph comes to allow us to build such reliable dictionary.

b) In this Oriented Graph, a vertex represents a word and each edge on the basis of this vertex is labeled with the probability that it might follow the word in the vertex. This probability is simply computed by dividing 1 by the total number of words having this vertex as basis multiply by the probability that the n + 1 words together are likely spam than legitimate words, where n + 1 is the number of previous words form the basis vertex to the actual visiting word. That means the arrow that points to them has this vertex as ancestor. It computes as follow:

$$P(x,y) = \frac{1}{\sum w_i} \times f(wi) \qquad (3)$$

Where the elements in equation (3) are defined as fellow:

- P(x, y) is the probability that a next word would be ‘y’ if the first word is ‘x’;
- $\sum w_i$ is the total number of word that might follows ‘x’;
- $f(wi)$ is the probability that the words from wo to wi are appeared likely in spam message than in legitimate message. This probability is computed during the learning process.

c) Next, sum up the whole probability of each sentence and compute its probability as in the case of natural Bayes Technique. Finally, make a decision based on the threshold of 95% or 80%.

B. *Dealing with rare words*

In the case a word has never been met during the learning phase, both the numerator and the denominator are equal to zero, both in the general formula, in the spamicity formula and in our formula. The software can decide to discard such words for which there is no information available. More generally, the words that were encountered only a few times during the learning phase cause a problem, because it would be an error to trust blindly the information they provide. A simple solution is to simply avoid taking such unreliable words into account as well.

C. *Sample size and data analysis*

Using a corpus from Ling Spam [9], we were able to test our application. The Ling Spam Corpus downloaded in that day of 17th March 2013 had 9640 legitimate emails and 1920 spam emails. We decided to use a sample size of 152 forms by 96 legitimate emails and 56 spams selected among the population of 9640 and 1920 respectively using a Simple Random Sample (SRS). These emails have been considered as our training set. A file name “spam.txt” and “good.txt” have been designated and filled in respectively with 96 legitimate emails and 56 spams emails.

After the process of learning with the aforementioned set of legitimate and spam emails. To test our application using the same corpus from Ling Spam, we selected randomly 10 spam and 10 legitimate emails. 10 of the legitimate emails have been classified as legitimate that is an accuracy of 100% and 9 of the 10 spam have been classified as spam with 1 as “may be spam” which leads to an accuracy of 90%. More other tests have been conducted in order to prove the accuracy of our model. The table below shows the recapitulation of our tests.

	Spams	*Goods*	*Cl. Spam*	*Cl. Goods*	*%*
Test 1	10	10	9	10	95%
Test 2	45	35	45	34	98.5%
Test 3	50	60	50	60	100%
Test 4	100	180	100	180	100%

Result of the test conducted upon our model

Based on this table of tests, we can conclude that our model attains 100% when the number of the emails increase and becomes stable with a threshold of 100 or 120 emails involved.

VIII. Conclusion

All through this research, we demonstrated the necessity of improving an existing technique for email classification which is Bayesian Technique. Next, we proposed an improved technique for spam classification based on Bayes. This improvement makes use of Data Structure called Graph in order to build a dictionary of words contained in the incoming email. Tests have been done and it has proved that, our proposed model perform approximately 99.9% of good classification. Therefore, one of the major contributions to knowledge is that, we derived a good algorithm for spam recognition which recognizes spam more accurately than the entire existing spam filtering based on Bayes in the world.

IX. FURTHER STUDIES

The accuracy and effectiveness in classification is reached at the expense of memory usage and processing time. Also, the time required for searching word with related others words in the same hierarchy increase as well as the words in the message increase. Therefore, one of the recommendations concern developing In addition, we also develop strategy for reducing the processing time and memory usage.

X. REFERENCES

[1] David Mertz, Analyzer, Gnosis Software, Inc., “Spam filtering techniques, six approaches to eliminating unwanted e-mail”.

[2] Vikas P. Deshpande, Robert F. Erbacher, and Chris Harris, “An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques”. Proceedings of the 2007 IEEE, United States Military Academy, West Point, NY 20 – 22. June 2007. **(Article in a conference proceedings).**

[3] Trevor Stone Department of Computer Science University of Colorado at Boulder Masters Comprehensive Exam, “Parameterization of Naive Bayes for Spam Filtering”, Fall 2003.

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

[5] Ion Androutsopouls, John Koutsia, Konstantinos V. Chandrion, George Palioura and Constantine D. Spyropoulos, “An Evaluation of Naïve Bayesian Anti-Spam Filtering” 11th European Conference on Machine Learning, Barcelona, Spain, pp. 9 – 17, 2000**. (Article in a conference proceedings).**

[6] Tich Phuoc Tran, Min Li, Dat Tran and Dam Duong Ton , “Spam Recognition using Linear Regression and Radial Basis Function Neural Network”, Pattern Recognition edited by Peng-Yeng Yin, In-Tech, intechweb.org, pp. 513-531, Octobre 2009.

[7] Bayesian Spam Filtering, http://en.wikipedia.org/wiki/Bayesian_spam_filtering last visited April 2nd.

[8] Abstract Data type Graph http://en.wikipedia.org/wiki/Graph_(abstract_data_type)

[9] Ling Spam Datasets, Project Dataset http://csmining.org last visited on March 17th 2013.

[10] Jingrui He, Bo Thiesson, Microsoft Research, Carnegie Mellon University, “Asymmetric Gradient Boosting with Application to Spam Filtering”

[11] Xavier Carreras and Luis Marquez, “Boosting Trees for Anti-Spam Email Filtering”, 2001. Universitat Politecnica de Catalunya (UPC).