Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks

Ann Nosseir [1], Khaled Nagati [1] and Islam Taj-Eddin [1]

[1] Faculty of Informatics and Computer Sciences
British University in Egypt (BUE)
El Shorouk City, Cairo, Egypt

Abstract

SPAM e-mails have a direct cost in terms of time, server storage space, network bandwidth consumptions and indirect costs to protect privacy and security breaches. Efforts have been done to create new filters techniques to block SPAM, however spammers have developed tactics to avoid these filters. A constant update to these techniques is required. This paper proposes a novel approach which is a characters-word-based technique. This approach uses a multi-neural networks classifier. Each neural network is trained based on a normalized weight obtained from the ASCII value of the word characters. Results of the experiment show high false positive and low true negative percentages.

Keywords: *electronic mail (E-mail); spam filters; spam detection; Artificial Neural Network; stemming process.*

1. Introduction

SPAM is defined as an unwanted of electronic message [25] posted blindly to thousands of recipients [12] also known as 'junk e-mails'. They are unsolicited mails sent in bulk (unsolicited bulk E-mail) with a hidden or forged identity of the sender, address, and header information [11][22]defines an electronic message as a "SPAM" if (A) the recipient's personal identity and context are irrelevant because the message is equally applicable to many other potential recipients; and (B) the recipient has not verifiably granted deliberate, explicit, and still-revocable permission for the mail to be sent in the information and message body.

The problem is that SPAM imposes direct cost in terms of time, money, and storage space and indirect cost to protect privacy and security breaches. Users are inconvenienced by the SPAM because of the time they spend to filter legitimate email from SPAM emails. These unproductive hours can be calculated "based on the number of SPAM emails users read. Users could consume up to 4.33 hours per day dealing with SPAM [6]." For an organization, that could be translated into money terms by calculating the salary per hour. We can incur the operations and helpdesk running costs caused by the SPAM troubles [6]. The cost of computer and network resources consumed. These SPAM messages eat up a tremendous amount of storage space and cause bandwidth losses. There are also indirect costs in managing SPAM and protecting information. These costs are related to criminal acts caused by spamming. For example, financial theft, identity theft, data and intellectual property theft, virus and other malware infection, child pornography, fraud, and deceptive marketing are usually caused by SPAM messages that point to links to collect personal information, open porn Websites, or download virus [6]. These costs are calculated depending on the security strategy the company or the people employ to protect the information or equipment.

Some countries such as Denmark, USA, and Canada have realized the economic impact of SPAM and have enacted different laws and legislations to protect businesses and individuals alike against SPAM [23]. For example, the USA CAN-SPAM Act is a law designated to control the assault of non-solicited pornography and marketing. It places restrictions and regulations to control spammers' activities. It prohibits spammers from harvesting e-mail addresses and creating Bot-nets [3].

Besides the legislative approach, there are different approaches to control SPAM. One approach is to check the message body, IP address, or domain names to filter the legitimate from SPAM messages [19][8][4][9][21][2] for example, black-list and white-list and heuristic approaches. Using white-list and black-list approach, the filter is a content-based technique that recognizes words or patterns of message contents. The e-mail message is analyzed against the lists of the one that matches the content in the black-list and blocked while the other are listed in the in white-list and become legitimate e-mails [7][20].

The heuristic approach examines the e-mail's content and compares it against thousands of pre-defined rules [24]. These rules are assigned a numerical score that weight the probability of the message being SPAM. Each received message is verified against the heuristic filtering rules. The score of the weight is then shared among users to filter the e-mails [15].

In spite of the effectiveness of the current SPAM filters techniques used to solve the SPAM problems, spammers bypass these kinds of techniques by periodically changing their practices and behaviors [10]. A continuous enhancement to the current techniques and developing new ones are important to control SPAM.

This paper examines related work in the area of Spam filters and gives a detailed description of our proposed solution, including procedures, results, and conclusions. We conclude with a discussion of the findings and some directions for future work.

2. Related Work

Most of the techniques applied to reduce SPAM email is the content-based filtering. These filters are Origin-Based Filters or Content Filtering. Both filters mainly classify the messages as SPAM by using different approaches.

Origin based filters methods use network information such as IP and the email address which are the most important pieces of network information available. The major types of origin-Based filters are Blacklists, Whitelists, and Challenge/Response systems [5].

Content filters read the text to examine its content. They are called Keyword-Based Filters. There are several popular content filters such as: Bayesian filters, Rule Based Filters, Support Vector Machines (SVM) and Artificial Neural Network (ANN) [5][19].

Bayesian Filters are the most well known commercial machine learning approaches. They calculate and use the probability of certain words/phrases occurring in the known examples (messages) to categorize new examples (messages) [19]. Support Vector Machine SVM has prompted significant research into applying SPAM filtering [19][1][16]. SVMs are kernel methods whose central idea is to embed the data representing the text documents into a vector space where linear algebra and geometry can be performed [19][16]. SVMs attempt to construct a linear separation between two classes in this vector space [19]. Rule-Based Filters use a set of rules on the words included in the entire message (Header, Subject, and Body) [17]. The limitation of Rule- Based filtering is a rule set which is very large and static and not fully effective. The spammers can easily defeat these filters by word obfuscation, for example, the word "Free" could be modified to be F*R*E*E so it will pass the filters [5][19].

Even with the efforts to reduce the SPAM, it is still considered to be a threat. MessageLabs Intelligence reports that in 2010, the average global SPAM rate for the year was 89.1%, an increase of 1.4% compared with 2009 [13]. One reason could be the increase in the number of Internet users or spammers learn the filters techniques and continue their criminal behavior. This drives research to update their filters techniques and approaches.

This work suggests a novel filter technique which is an intelligent content-based filter. It investigates the feasibility of creating a multi-neural network filter at the level of the word and combinations of the letters.

3. Proposed Solution

This research looked into junk e-emails of several academics within a department and generated three lists of three, four and five characters words. These lists had two categories of words, bad and good, extracted from the junk mails. To produce these lists, the message content has been processed- preprocessing phase- in three steps. In the first step the stop-word removal, a list of stop words such as articles (e.g. "a", "an" and "the"), prepositions (e.g. "with" or "beside") and conjunctions (e.g. "and", "or" or "for") have been generated and compared against the message content to get rid of words that are mapped to the list [24]. In the second step, the stop-word removal, the work has generated another list of noise. Noises such as misspelling, misplaced space or embedding special characters are extracted. For an instance, the word Viagra could be written as "V1agra", "V|iagra" or Free into"fr33"[23]. The message content has been compared against this list in order to remove the message noise.

In the third step, the words in message went into a stemming process which reduces words into their basic form. The process strips the plural from nouns (e.g. "apples" to "apple"), the suffixes from verbs (e.g. "measuring" to "measure") or other affixes. Originally proposed by Porter in 1980 [14], it defines stemming as a process for removing the commoner morphological and in-flexional endings from words in English. A set of rules is applied iteratively to transform words to their roots or stems.

Afterwards, the words have been classified by its length and by being good or bad words. The bad words are commonly used in the junk e-mail messages (See Table 1).

Table 1: Example of Good and Bad word in each list

List	*Good word*	*Bad word*
3 characters word list	AIR	SEX
4 characters word list	POET	SHIT
5 characters word list	CLAIM	STRIP

Using these lists, the research managed to train multi-neural networks on the bad and good words and test the results. The multi-networks identify the bad and good words in the message. If the message doesn't have bad words, it is classified as a good message i.e., HAMS. Otherwise, a weight is added to the bad words according to its category. We have categorized the words into advertisements, commercial, financial, and pornography category. Advertisements and commercial category gets value 1; financial and pornography categories get value 2. The value of the category could be changed by the users based on their requirements. Some users, for example, get annoyed by pornographic messages than commercials ones. Therefore, they can modify the weight of each category. The identified bad words by the Multi- networks are classified under their categories and then multiplied by their categories' weights. A decision function based on the calculated value of the message could be used to decide whether the e- mail message is a SPAM or not (See Figure 1).

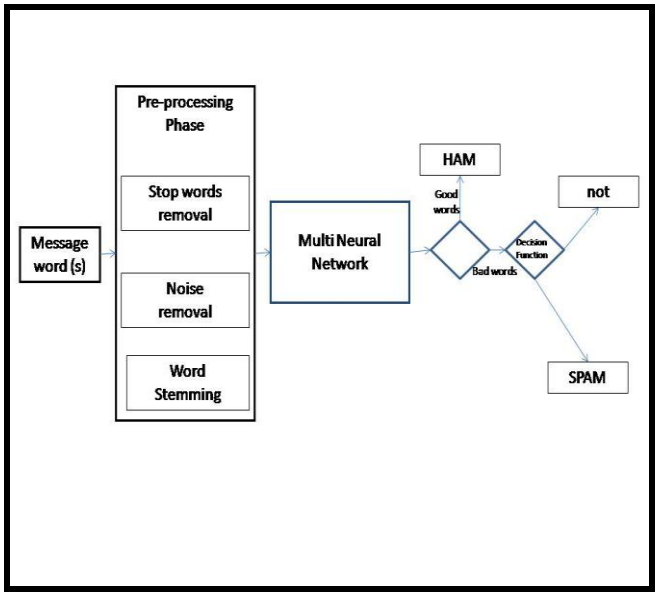

Figure 1. The proposed approach

4. Multi-Neural Networks Classifier

After the pre-processing phase, the data collected are processed. Each character in the word is converted into its ACSII values then, gets a normalized weight which is a value ranging from 0 to 1. We have used the following equation to get the normalized value for each character (See Equation 1).

$$Wi = \frac{(ACSII\,(Char_i) - ACSII\,(Char_{Min}))}{ACSII\,(Char_{Max})} \quad (1)$$

W_i : the normalized weight of the character
Char : character
I : the character position (from 1- A…26-Z)
ACSII : ACSII value of the character Classifier

We created a network for each list. In other words, we had three characters, four characters, and 5 characters neural networks. All networks are back-propagation neural network. The three characters neural network has three-layers with 3 neurons in an input layer and 3 neurons in a hidden layer and 2 neurons in an output layer. The four characters neural network has three-layers with 4 neurons in an input layer and 4 neurons in a hidden layer and 2 neurons in an output layer. The five characters neural network has three-layers with 5 neurons in an input layer and 5 neurons in a hidden layer and 2 neurons in an output layer. The output neurons of all networks have values 1 and 0 representing bad and good words (see Figure 2).

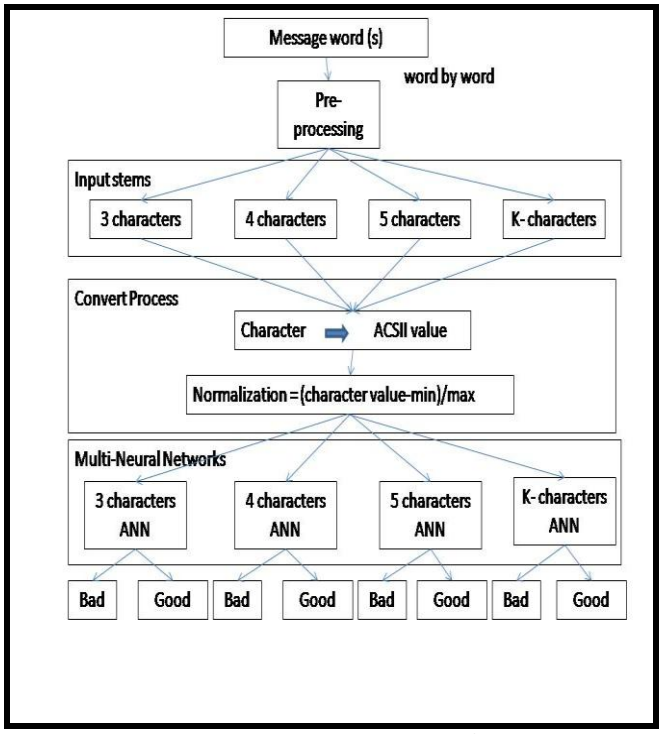

Figure 2. Multi-neural networks classifier

5. Experiments

5.1 Data Test

To train the multiple networks, there were three sets of sample data with length of 3, 4, and 5 characters. Each set contains 40 words divided into bad and good words where 50% of the words assigned "bad" and other 50% assigned "good"(see Table 1). Each sub network is trained on 80% of the data set and 20% of this data was used to test and validate the network results.

5.2 Results

The sensitivity and specificity technique was used to analyze the results. Tables 2, 3, and 4 show the results in confusion matrices. Type I false positive and Type II true negative results were as follow 0.131364 and 0.999962, 0.0003 and 0.7953, 0.0015 and 0.9990 for the three, four and five character neural network respectively.

The low false positive and high true negative results show the potential of the approach. Techniques such as the black list filter and white list filters treat the word as a black box, i.e., without considering the length of the word [12][20]. In our approach, we analyzed the relationship between the length of the word and its potential acceptance or rejection. We trained each sub-network on a set of words with specific word length. The sub-networks performance could improve the filter performance.

Table 2: Three characters neural network t

Three characters	***True***	***False***
Positive (bad word)	0.868637	0.131364 (Type I) blocks-legitimate users
Negative (good word)	0.999962 (Type II)	0.000038

Table 3: Four characters neural network

Four characters	***True***	***False***
Positive (bad word)	0.997	0.0003 (Type I)
Negative (good word)	0.7953 (Type II)	0.2047

Table 4: Five characters neural network

Five characters	***True***	***False***
Positive (bad word)	0.9985	0.0015 (Type I)
Negative (good word)	0.9990 (Type II)	0.0011

6. Conclusions and Further Works

Our novel approach uses a multi-neural networks classifier to identify bad and good words in the textual content of an email. Words in the message are preprocessed before using the multi-neural networks classifier. The word goes through stop words and noise removal steps then stemming process step to extract the word root or stem. The experiment shows positive results.

In the future, we would like to use a large sample database to thoroughly test the performance of the classifier. Additionally, consider a feedback from the users about the bad or good and retrain multi-neural networks classifier could improve results. To enhance the decision fusion quality of whether the email is SPAM or not, the word category weight could be personalized to adapt with users requirements. Finally, incorporating this classifier as a step before implement a content filter, such as Black and White lists and Bayesian Filters [12][20][18].

References

[1] A. Kolcz, J. Alspector, "SVM-based filtering of email spam with content-specific misclassification costs", TextDM'2001, IEEE ICDM-2001 Workshop on Text Mining, pp. 123-130, 2001.

[2] A. Wiehes, "Comparing Anti Spam Methods", Master Thesis, Master of Science in Information Security, Department of Computer Science and Media Technology, Gjøvik University College, 2005.

[3] A. Wosotowsky, E. Winkler, "Spam Report McAfee Labs Discovers and Discusses Key Spam Trends" 2009, Retrieved 10th Feb2010 http://www.mcafee.com/us/local_content/reports/7736rpt_spam_1209.pdf

[4] C. Gulyás, "Creation of a Bayesian network- based meta spam filter, using the analysis of different spam filters", Master Thesis, Budapest, 16th May 2006.

[5] D. Cook , "Catching Spam before it arrives : Domain Specific Dynamic Blacklists", Australian Computer Society, ACM , 2006.

[6] Ferris Reseacrh, "Spam, Spammers, and Spam Control A White Paper", March 2009, Retrieved 8th Feb 2010 http://apac.trendmicro.com/imperia/md/content/us/pdf/products/enterprise/interscanmessagingsecuritysuite/wp01_antispamferris_090311us.pdf

[7] G. Dalkilic, D. Sipahi, M.H. Ozcanhan, "A simple yet effective spam blocking method", Proceedings of the 2nd international conference on Security of information and networks, pp. 179-185, 2009.

[8] G.R. Weinberg, "A System Analysis Of The Spam Problem", Master Thesis, Submitted To The Engineering Systems Division in Partial Fulfillment Of The Requirements For The Degree Of Master Of Science In Technology, The Massachusetts Institute Of Technology, 2005.

[9] http://spam.abuse.net

[10] J. María, G. Cajigas, E. Puertas, "Content Based SMS Spam Filtering", DocEng'06, ACM , Amsterdam, The Netherlands, 1-59593-515-0/06/0010, October 10–13, 2006.

[11] J. Wu, T. Deng, "Research in Anti-Spam Method Based on Bayesian Filtering", IEEE Pacific-Asia Workshop on

Computational Intelligence and Industrial Application, pp. 887 – 891, 2008.
[12] L. Lazzari, M. Mari, A. Poggi, "A collaborative and multi agent approach to e-mail filtering", IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'05), pp. 238-241, 2005.
[13] MessageLabs intelligence, Annual Security Report (2009) Retrieved 8th Feb 2010 http://www.messagelabs.co.uk/intelligence.aspx
[14] M.F. Porter, "An Algorithm for suffix Stripping", 1980. Available at http://tartarus.org/~martin/PorterStemmer/def.txt
[15] M. Xie, H. Yin, H. Wang, "An Effective defense against e-mail spam laundering", ACM. Retrieved 2006, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.9653&rep=rep1&type=pdf
[16] N. Cristiani, B. Scholkopf, "Support vector machines and kernel methods", The new generation of learning machines, AI Magazine 23(3), 31-41, 2002.
[17] P. Graham, "better Bayesian Filtering", spam conference, 2003.
[18] S. Delany, "Using Case-Based Reasoning for Spam Filtering", PhD Thesis, A thesis submitted to the Dublin Institute of Technology in fulfilment of the requirements for the degree of Doctor of Philosophy School of Computing, Dublin Institute of Technology ,2006.
[19] S.J. Delany, "Using Case-Based Reasoning for Spam Filtering", A PhD Thesis submitted to the Dublin Institute of Technology in fulfillment of the requirements for the degree of Doctor of Philosophy School of Computing, Dublin Institute of Technology , 2006.
[20] S. Heron, "Technologies for spam detection", Network Security, pp. 11-15, Jan 2009.
[21] S. Hershkop, "Behavior-based Email Analysis with Application to Spam Detection", PhD Thesis, Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences, Columbia University, 2006
[22] Spamhaus, "Definition of Spam", Retrieved 8th Feb 2010 http://www.spamhaus.org/definition.html
[23] T.S Guzella, W.M. Caminhas, "A review of machine learning approaches to spam filtering", Expert system with Application, 36:10206- 10222, 2009.
[24] T. Subramaniam, H.A. Jalab, A.Y. Taqa, "Overview of textual anti-spam filtering techniques", International Journal of the Physical Sciences, Vol. 5(12), pp. 1869-1882, 4 October, 2010, Available online at http://www.academicjournals.org/IJPS\
[25] V.N. Vapnik, H. Druck, D. Wu, "Support Vector Machines for Spam Categorization", IEEE Transactions On Neural Networks, 10(5): 1048-1054, 1999.

Ann Nosseir Ph.D. in Computer Science, University of Strathclyde; currently is a Lecturer at The Faculty of Informatics and Computer Science, The British University in Egypt.

Khaled Nagati Ph.D. in Computer Science, Joint supervision: American University in Cairo (AUC) and Cairo University; currently is an Associate Professor at The Faculty of Informatics and Computer Science, The British University in Egypt.

Islam Taj-Eddin Ph.D. in computer Science, The Graduate School and University Center, The City University of New York; currently is a Lecturer at The Faculty of Informatics and Computer Science, The British University in Egypt.