IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 2, September 2016)

A Review of Spam Detection using Machine Learning

Varsha Malik
M. Tech. Scholar
Dept. of Computer Science and Engineering
BSA College of Engineering and Technology,
Mathura (India)
varsha.mlk@gmail.com

Sanjay Kumar
Asst. Professor
Dept. of Computer Science and Engineering
BSA College of Engineering and Technology,
Mathura (India)
sanjaysingh.chauhan@bsacet.org

Abstract –**The computerization of communications has increased the velocity of trade and greatly enriched the content. Emails, franchisees in emails (emails contraction) are increasingly used by individuals and more by businesses (70 billion emails per day). But as for traditional mail, users very quickly deal with unwanted email, or spam, and mostly quite undesirable. To counter this flood of trash (more than 50% of all email) and not lose emails that we are actually intended, only one viable solution: automate the detection and destruction of this type of digital pollution with the risk that a document or misfiled.**
This paper aims to present the progress of spam detection techniques.

Keywords –**Classification, e-mail, SPAM.**

I. Introduction

Unwanted email, or spam (trash contraction and email), represent a very large share of global email traffic. According to various analyses, the tens of billion emails over the network every day, almost 50% are spam and this number is increasing. Moreover, many of these are spam propagation vectors for some viruses and worms that bring digital security of data stored on our computers to the test.

Although difficult to quantify, the cost of digital pollution represent more than $200 billion per year in the World for Users (lost productivity, connection cost, detection software, etc.).

We now understand the importance of application development to fight effectively against this form of pollution.

Difference between Desirable and Undesirable Email

By definition, an email is undesirable when one hand it was not solicited and / or content that are either relevant, desired or deemed worthwhile by the user and thus a pollution healthy messages and finished so directly into the trash. But how to judge the relevance of the content of an email? And worse, how to determine that a message is of interest to the user? This touches the heart of email classification problem. By using considerations inherent to the user and the context of use of its email service (work, personal, etc.), it immediately loses the ability to sort messages on strict criteria based on simple characteristics and free from ambiguity.

Failing to understand the content of an email, and more importantly, to be able to make a judgment in lieu of the user, it can, in most cases, simply to analyse the structure, origin, destination and edge effects an email to detect any deviations from a standard derived from the behaviour and context of the user.

To detect deviations and suspicious characteristics, current techniques will rely mainly on the analysis of user behaviour and thus on statistical data to perform probabilistic classification. So we can never have absolute certainty about the classification of an email and we always keep to automatically delete junk mail judge.

II. Solutions Against SPAM

There are two types of email classification software. One is directly from the user and is generally based on their behaviour to classify emails and the other is located directly at the entry point of emails among email service providers, and that will usually (but not only) rely on the signing of emails and associated with the volume of emails passing through similar service to detect abnormal and massive shipments.

Techniques Used:

1. Foreword

To judge the performance of a classifier, the following concepts are used:

- True Positive ratio of class *A* elements have been labeled *A* through the classifier.
- False Positive Ratio: class *A* elements have been labeled *B* by the classifier.
- True Negative Ratio: Class *B* elements have been labeled *B* by the classifier.

- False Negative Ratio: Class *B* elements have been labeled *A* through the classifier.

Apart from the overall correct classification ratio (True Positive + True Negative), the False Positive ratio which will interest us in highest place because losing an important email due to misclassification can have disastrous consequences. The user, depending on the degree of risk of misclassification, must be able to control these risks and be able to adapt classification policy.

Here is a list of techniques commonly used by anti-spam solutions:

- Confrontation signature against a spam database
- Reputation (social networking)
- Probabilistic classifiers: Naive Bayes, Quadratic discriminant analysis
- Support Vector Machine
- Neural Network
- Data-mining
- Whitelists / black
- Clustering
- Genetic Algorithm

2. *Probabilistic Classifiers*

To perform a probabilistic classification must first identify the characteristics of emails that differentiate them. Based on selected characteristics, and each email, we will create a feature vector which will allow to establish a classification.

Here is a list of characteristics of potentially discriminatory emails and which will train classifiers:

- Length of the message and the subject.
- Number and type of attachments
- Presence of HTML, scripts and hyperlinks
- Presence of image (s)
- Sender: is it known by the user?
- Is the area of service sends blacklisted? Is it permitted to perform this type of shipment?
- Recipient (s): unique, mailing list, hidden copy, blind copy.

From the frequencies obtained for each characteristic, the selected classifier will produce a probability of belonging to each of the classes as possible and the most likely possibility is held by:

Advantage (s): good performance.

Inconvenient (s): Performance depends on the quality of training; requires continuous training to make new forms of spam [10], [13].

3. *Data-Mining*

The other major classification technique is that of data-mining algorithms supervised based decision tree such as C4.5, SLIQ or CART. Emails previously classified by the user is cut according to relevant characteristics and inserted into the database. From these records, the inference algorithm will generate a decision tree that may need to be converted set of rules and directly integrated with email clients supporting.

The quality of the classification depends on the quality of data preparation and recordings selected for the training base:

Advantage (s): effective as long as the training base is good.

Inconvenient (s): difficult to implement; requires the construction of a new decision tree to cope with new types of spam [16].

4. *Artificial Immune System*

To fight effectively against spam, the researchers made the parallel between spam and pathogens that are fought by the human immune system. From a gene library, the system will randomly generate antibodies and create the corresponding cell in order to be able to detect any foreign bodies entering the system. The lymphocytes are then trained on a database of emails and spam, inefficient ones are removed, and an expiration date is assigned to each cell based on its performance. In each message filtered by a cell, the expiry date is increased. Expired lymphocytes die and are replaced by others that restart a cycle.

Advantage (s): effective; lightweight (only 200 antibody run simultaneously); possible use of emergency vaccine [12].

5. *Neural Network*

Neural networks allow, after learning, to reproduce a form of human reasoning. The characteristics of emails and their contents used to adjust the synaptic coefficients of the neural network during the learning phase. Learning is from a collection of emails pre-sorted by the user and can optionally be incremental to be best suited as possible to new forms of spam that may appear. After learning made, the neural network works as a highly effective conventional anti-spam system according to the case.

Like any classifiers, misclassification risk (False Positive) is real but can be controlled by adjusting the neural network of the sensitivity threshold (at the expense of the False Negative).

Advantage (s): Sets the misclassification rate (False Positive) by adjusting the sensitivity threshold; fast.

Inconvenient (s) requires a long training; must be regularly trained to deal with new forms of spam [11].

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 5, Issue 2, September 2016)

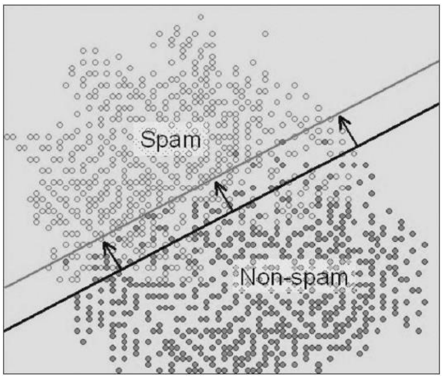

Figure 1: Setting the neural network of the sensitivity threshold

6. *Message Signature:*

To fight against mass emails, the anti-spam system needs to position itself at the email sending service to have a global vision of emails over the network and be able to detect the massive shipments characteristics of spammers.

However, to make the massive shipments more difficult to detect spammers insert or remove random sequences in emails so that each campaign message is unique. Cannot be based solely on a checksum to identify identical emails detection techniques must be based on another form of signing less sensitive to the insertion/removal of terms. This is the case of I-Match algorithm.

The I-Match algorithm is based on all of the unique words and email on a lexicon previously established to produce the message signature. This signature is then associated with a single cluster which allows to deduce the message class.

Advantage (s): detection of massive shipments; insensitive to random changes in the body of email.

Inconvenient (s) must be set up by the email service provider; must be used with complementary technology; real risk of misclassification (false positive) [8].

7. *Reputation of the Issuer (Social Networking)*

The most radical technique to sort the emails is still that white lists (white-list) which is to manually set or semi-automatically a contact list that is trusted and known that the emails are "valid". In practice, the emails sent by people you trust are classified in the Inbox and the other is sent to a subdirectory for junk mail.

But this technique suffers from two major problems. On the one hand, the user must himself keep the white list which may, in some cases, represent an amount of significant work. Secondly, and this is the biggest drawback of this technique is what is based solely on known contacts while many emails can come from individuals or organizations (known and unknown) not yet listed in whitelist and thus end up mixed in the subdirectory reserved for junk mail.

To overcome this problem, one technique is to use a reputation network [4] will note (between 0 and 10) each issuer based on its knowledge network which will allow to award (for each user) an index of confidence and thus better classify emails. On the other hand, confidence attributed to emails sent by unknown transmitters can be inferred for some issuers are known by a person or persons of the social network of the user.

Although significantly improved quality classification, this technique does not solve the problem completely unknown transmitters for each social network has only a finite number of individuals. Also, always keep a white list and record each of his contacts for the system to function optimally. So this is an interesting but insufficient in itself to technical use other classification techniques complement.

Advantage (s): to fight indirectly against the misclassification rate (False Positive).

Inconvenient (s): difficult to maintain; requires a large social network; solves the problem of unknown transmitters [4].

8. *Authentication of the Sender (Turing Test):*

One of the most effective techniques is the issuer authentication that is based on a simple fact: spam is sent automatically by computers. Based on this fact inhering to mass emailing, just automatically identify the sender as actually being an individual by asking a question to the issuer that only a human can answer.

For example, the system can send a captcha (an image containing distorted and noisy enough characters to seriously complicate the task with OCR) to the email sender (only the first time) and ask him to bring the answer to the question (copy the text written in the image or make a mathematical operation).

Although radical and effective, this solution suffers from several problems caused for certain applications. First, the user will have to set up a whitelist for organizations that send messages automatically (administrative site, e-commerce, etc.) which can quickly become laborious. Then it's a constraint method to the message sender. In a context of owners where the volume is not very important not particularly annoying but once during the daily volume of email exceeds a certain threshold - which is rapidly becoming the case in the workplace - it quickly became expensive for users to answer authentication questions. Finally, this technique requires the use of an external control

platform, storage authenticated and routing email users.
Advantage (s): effective; Quick to set up. Inconvenient (s): difficult to maintain (white lists); restrictive for the issuer.

9. Other Techniques

RBL (Real-time Black-hole List): these huge common database containing the list of all servers known to be used for spamming. If the sending server is listed in a RBL, then this is that it is spam.
SPF (Sender Policy Framework): it is verified that, in the DNS zone of the domain, the server is allowed to make shipments.
Gray List: This technique is based on the standards described in RFC 2821, which stipulate that a mail receiving server, in case of unavailability, must return error code 421 to the transmission server that will have to wait a sometime before re-transmitting the email. Spammers, to save time, return emails much earlier than the minimum time. Just then pass only emails that are sent after the minimum waiting time. This technique is very easy to set up by the receiving server administrator is very effective but adds a lag time for the recipient.

III. Solutions for E-Mail Servers

SpamAssassin: Open Source system, free GPL licensed, reputable and effective but difficult to configure and slower than other business systems as MSwitch Anti-Spam.
M-Switch Anti-Spam: paying system for mail server. One of the most effective according to several studies (including that of Isode), especially regarding the False
Positive.
SpamGuru: System developed by IBM [15] on the basic TEIRESIAS [14] algorithm (detection of gene sequences) which are added several filters and other heuristics. The best anti-spam market (98% spam detection and only 0.1% False Positive).

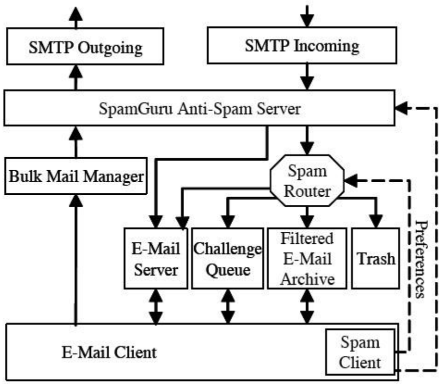

Figure 2: Architecture of SpamGuru

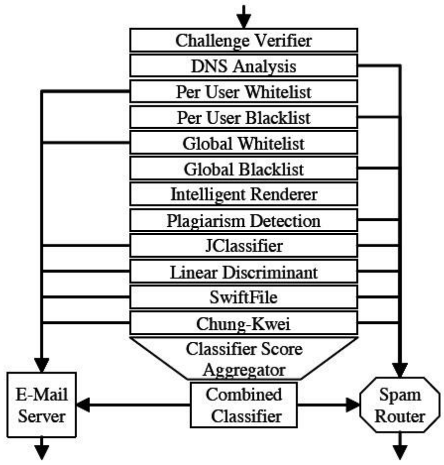

Figure 3: SpamGuru pipeline

IV. Solutions for the End User

Among these solutions on the market are:

- Bogofilter: GPL, free, cross-platform, Bayesian filter.
- Vade Retro: high performance and Free French solution (without updating).
- SPAMfighter: only works with Outlook®, available in free or paid version.
- SpamPal: works only on Windows®, Free RBL.
- Spamihilator: works only on Windows®, Free Bayesian filter.

V. Literature Review

The existing work undergo an implementation on detection of malicious URL in Email by Dhanalakshmi R and Chellapan C. considered Age of domain, Host based features, Lexical features and Page rank for analysis of URL to classify into malicious URL and legitimate URL. They have used Bayesian classifier to improve the accuracy by reduced feature sets and considered phishtank dataset, the work was restricted to URL in Email only [17].
Sahami et al. presented a spam classification method using a Bayesian approach. A Bayesian classifier is statistical classifier works on independence computation of probability. They have considered content of e-mail with features of domain, and shown that accuracy can be increased [18].
V Christina et al., shown that the need of effective spam filters increases. He discussed spam and spam filtering methods and their correlated problems [19].
Sadeghian A. et al. presented spam detection based on interval type-2 fuzzy sets. This system gives user more control on categories of spam and permits the personalization of the spam filter [20].

Congfu Xu et al. derived a feature extraction on Base64 encoding of image with n-gram technique. Effectiveness and efficiency in detecting spam images are shown by these features from legitimate images by training a SVM. Experimental results shows that it has prominent performance for classification of spam image in terms of accuracy, precision, and recall [21].

Man Qi et al. explored two main semantic methods: Bayesian algorithms and Support Vector Machine (SVM). Recent spam filters are discussed in this paper for determining spam messages which utilize semantic analysis information [22].

Zhan Chuan, LV Xian-liang presented an application to Anti-Spam Email using a new improved Bayesian-based email filter. They have used vector weights for representing word frequency and adopted attribute selection based on word entropy and deduce its corresponding formula .It is proved that their filter improves total performances apparently [23].

Holly Esquivel et al. focused on the pre-acceptance altering mechanism IP reputation. They first classify SMTP senders into three main categories: legitimate servers, end-hosts, and spam gangs, and empirically study the limits of effectiveness regarding IP reputation filtering for each category [24].

Georgios Paliouras et al., presented Learning to Filter Spam E-Mail. They investigated the performance of two machine learning algorithm in context of anti-spam filtering by comparison of a NaIve Bayesian and a Memory-Based Approach. They have determined the performance on publicly available corpus for naive bayes. Also they have compared the performance of the Naive Bayesian filter to an alternative memory based learning approach so that in both methods accuracy has improved for spam filtering and keyword based filter are used widely for email [25].

Gray Robinson proposed computation of probability of spam mail and legitimate mail [26].

VI. Future Developments

The spam detection techniques based on genes, antibodies or genetic algorithms are now highly effective and very fashionable but new advances in temporal path detection [6] may restore the attractiveness the probabilistic techniques as well as data mining algorithms.

On the other hand, advances in the field of text mining could one day open a new path in the great family of email classification techniques.

VII. Conclusion

Whatever the technique we have seen that it is not possible to obtain a correct automatic classification to 100%. However, regardless of the technique used, it is seen that the solutions reach a correct classification rate and especially the False Positive rate quite correct or even well-priced for some.

But anyway, do not lose because the choice of technique to be used and the expected performance levels are strongly related to the context of use and the requirements associated with it. For a company, it is not acceptable to lose (by misclassification) a valid email while for a particular this can be much less severe. According to his expectations, so we choose to favour a False Positive rates low or high overall rate of correct classification. Similarly, the choice of solution must take into account the volume of emails to deal with and the sacrifices we are willing to concede.

Finally, we must not forget that it is often possible - and preferable - to combine different techniques to achieve the expected performance.

Reference

[1] David A. Bader and Ashfaq A. Khokhar, editors. Proceedings of the ISCA 17th International Conference on Parallel and Distributed Computing Systems, September 15-17, 2004, The Canterbury Hotel, San Francisco, California, USA. ISCA, 2004.

[2] Richard Clayton. Stopping spam by extrusion detection. In CEAS, 2004.

[3] Ernesto Damiani, Sabrina De Capitani di Vimercati, Stefano Paraboschi, and Pierangela Samarati. An open digest-based technique for spam detection. In Bader and Khokhar, pages 559–564.

[4] Jennifer Golbeck and James A. Hendler. Reputation network analysis for email filtering. In CEAS, 2004.

[5] Christian Jacob, Marcin L. Pilat, Peter J. Bentley, and Jonathan Timmis, editors. Artificial Immune Systems, 4th International Conference, ICARIS, 2005, Banff, Alberta, Canada, August 14-17, 2005 , Proceedings, volume 3627 of Lecture Notes in Computer Science. Springer, 2005.

[6] Svetlana Kiritchenko, Stan Matwin, and Suhayya Abu-Hakima. Email classification with temporal features. In Klopotek et al. [7], pages 523–533.

[7] Mieczyslaw A. Klopotek, Slawomir T. Wierzchon, and Krzysztof Trojanowski, editors. Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004, Advances in Soft Computing. Springer, 2004.

[8] Aleksander Kolcz, Abdur Chowdhury, and Joshua Alspector. The impact of feature selection on signature-driven spam detection. In CEAS, 2004.

[9] Vijay Krishnan and Rashmi Raj. Web spam detection with anti-trust rank. In AIRWeb, pages 37–40, 2006.

[10] Steve Martin, Blaine Nelson, Anil Sewani, Karl Chen, and Anthony D. Joseph. Analyzing behavioral features for email classification. In CEAS, 2005.

[11] Chris Miller. Neural network-based antispam heuristics. In Symantec, white paper, 2003.

[12] Terri Oda and Tony White. Immunity from spam: An analysis of an artificial immune system for junk email detection. In Jacob et al. [5], pages 276–289.

[13] Jefferson Provost. Naive-bayes vs. rule-learning in classification of email. Technical Report AI-TR-99284, The University of Texas at Austin, Department of Computer Sciences, 1999.

[14] Isidore Rigoutsos and Aris Floratos. Combinatorial pattern discovery in biological sequences: The teiresias algorithm [published erratum appears in bioinformatics 1998;14(2) : 229]. Bioinformatics, 14(1) :55–67, 1998.

[15] Richard Segal, Jason Crawford, Jeffrey O. Kephart, and Barry Leiba. Spamguru : An enterprise anti-spam filtering system. In CEAS, 2004.

[16] Ellen Spertus. Smokey : Automatic recognition of hostile messages. In AAAI/IAAI, pages 1058–1065 , 1997.

[17] Dhanalakshmi Ranganayakulu and Chellappan C., "Detecting malicious URLs in E-Mail - An implementation", AASRl Conference on intelligent Systems and Control, Vol. 4 ,2013,pg. 125-131

[18] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail", AAAI Tech. Rep.WS-98-05. pp. 55-62, 1998.

[19] V Christina., "A study on email spam filtering techniques", International Journal of Computer Applications, Vol. 12- No.1, 2010.

[20] Sadeghian, A and Ariaeinejad, R., "Spam detection system: A new approach based on interval type-2 fuzzy sets", iEEE CCECE -000379, 2011.

[21] Congfu Xu, Yafang Chen, Kevin Chiew, "An approach to image spam filtering based on base64 encoding and N-Gram feature extraction", IEEE international Conference on Tools with Artificial intelligence, DOI I0.1109/ICTAl.2010.31,2010.

[22] Man Qi, Mousoli, R,"Semantic analysis for spam filtering", international Conference on Fuzzy Systems and Knowledge Discovery, VoI.6, Pg. 2914-2917, 2010.

[23] Zhan Chuan, LU Xian-Iiang, ZHOU Xu, HOU Meng-shu, "An Improved Bayesian with Application to Anti-Spam Email ", Journal of Electronic Science and Technology of China, Mar. 2005, Vol.3 No.1.

[24] Holly Esquivel and Aditya Akella, "On the effectiveness of I P reputation for spam filtering", IEEE international Conference on Communication Systems and Networks, DOl: I 0.11 09ICOMSNETS.20 I 0.5431981, Pg.I-10, 2010.

[25] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memorybased approach", Proceedings of the Workshop on Machine Learning and Textual information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), pages 1-13,2000.

[26] G. Robinson., "A statistical approach to the spam problem", October 2014.