A Hybrid Approach for Spam Filtering using Support Vector Machine and Artificial Immune System

Kunal Jain
Student IEEE Member, Dept of CEA
National Institute of Technical
Teachers' Training and Research
Bhopal, India
kunaljain045@gmail.com

Sanjay Agrawal
Professor, Dept of CEA
National Institute of Technical
Teachers' Training and Research
Bhopal, India
sagrawal@nitttrbpl.ac.in

***Abstract*— Internet is a very powerful tool for information sharing; it provides email, chat, and audio/video talk for communication. All these email are widely used for official and non official communication because it is freely available to users and it also provides file transfer up to some limit. Hence use of Email (Electronic mail) services increasing rapidly due to higher dependency of organizations and individuals. Whenever Emails are sent/received unnecessarily then it is known as spam mail (Unsolicited Bulk Email). Spam Email is major issue for internet community because it causes wastage of resources and also pollutes our environment. Hence spam filtering is essential task. There are many existing techniques and algorithms available which focuses on individual parameters of the malicious content. Many times effectiveness of filtering algorithm gets significantly decreased, whenever spammer attacks on limitations of individual filtering mechanism. In this paper we have introduced an approach which includes advantages of Support Vector Machine and Artificial Immune System. In this approach we are trying to combine various positive properties of these filtering techniques at different level by deploying them in a hybrid approach. We have also discussed shortcomings of traditional spam filtering techniques in comparison of our proposed work.**

Keywords*— *Spam; Legitimate; Email (electronic mail); Bandwidth; similarity coefficient

I. INTRODUCTION

Email spam is generally defined as "unsolicited bulk mail". These emails are sent in bulk to large number of recipients [1]. Generally email spam messages contain advertising contents like credit card, home shopping, real state to local restaurant and everything. Because of its unwanted receiving, it is also referred as unsolicited or junk email. Everyday spammers sent thousands of spam emails. Mostly these spam email tries to sell something over the Internet. Most of the recipients delete these messages even without reading it. Very few people open and read these types of emails and among them very few people buy something to be sold or respond to these emails. These few people who respond or buy something give profit to spammers. Because spam email need very low cost to send hence it gives high benefit to sender.

The main property of spam email is, it is sent to varied recipients and its contents are advertizing type [2]. Many times by clicking on link associated with spam mail, it sends on phishing websites or websites which contains virus. Spam email may contain malware/virus as scripts or completely different viable file attachments, which is harmful for security of the system. Various times spam email demand for confidential data, and it leads to hacking or online frauds. Now it is clear that spam emails are unwanted interrupt for email users. An email is said to be spam if [1] -

- The recipient personally finds that the contents are irrelevant because the email is sent to multiple recipients and/or
- The sender doesn't have permission to send email to relevant recipient.

At major level spam emails are classified in six types given in Fig 1 [3]. With growth of Internet numbers of email users are increasing. Email statistics report given by Radicati group says, it is estimated that 1.1 billion email accounts will be created till 2017 in the entire world. There are 3.8 billion email accounts present in 2013 and it will be grown up to 4.9 billion in 2017 worldwide. Because of this increment a huge amount of spam email will be created. Some other numerical facts about spam email taken from various sources can be analyzed in TABLE 1.

All the numerical facts and figure are terrible and says about the future problems associated with spam email. To overcome this problem spam filtering is a solution. Generally spam filtering is a mechanism where we try to avoid spam email to arrive in user's Inbox. There are two major categories of spam filtering techniques exist-

- Content Based Spam Filtering and
- Meta data Based Spam Filtering

Both of these techniques have their own advantages and disadvantages. Detailed discussion on these techniques will be given in next section of the paper.

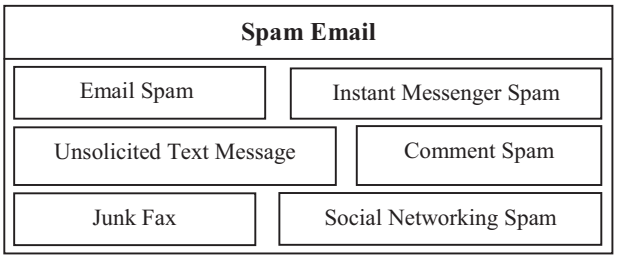

Fig 1. Classification of spam email at major level

TABLE I. STATICS ASSOCIATED WITH SPAM EMAIL [4]

Total Worldwide Emails Sent/Received Per Day in the year 2013	**182.90 Billion**
Total Spam Emails Sent/Received Per Day in the year 2013	**146.32 Billion**
Total amount of Spam received per person yearly (Approx)	**3500**
Estimated total cost for Spam email in 2013 (Global)	**200 Billion $ (Approx)**
Estimated increment in Spam email in 2017 (Approx)	**165.28 Billion (12.95%)**
Estimated global Spam cost in 2017	**250 Billion $ (Approx)**
Users who replied to Spam email in 2013 (Approx)	**30%**
Users who did business transaction in response to Spam email in 2013	**9%**
Estimated wasted corporate time per Spam Email in 2013	**30 Seconds**
Estimated CO_2 emission per day by Spam email in 2013	**43896 (Metric Ton)**

In this paper we review existing spam prevention techniques and proposing a new concept for spam filtering which includes hybrid strategy. These hybrid strategies have qualities of traditional spam filtering techniques together and will give better result than existing one. First section of this paper provides wide background knowledge about spam and spam filtering techniques. The entire organization of this paper is as follows: In second section, we have given an overview for existing spam filtering techniques with their disadvantages. And in third section; we discussed our proposed model for effective spam filtering with its advantages over traditional system. In section IV of this paper, we have discussed experimental setup for evaluating the proposed work. Section V has result analysis and in section VI we concluded the entire work with future research direction.

II. RELATED WORK

There are many solutions proposed in the recent past to merge different spam filtering approaches to find a single filter that has been sufficiently effective. Muthukaruppan Annamalai *et al* [5] captured the pleasing qualities of Naive Bayes and Decision Tree based classifiers in a hybrid spam filtering approach. This algorithm provides excellent learners for domains with a huge number of reliant attributes. They also described that huge number of attributes creates the domain not suitable for decision trees and naïve bayes will not perform as desired because of lack of independence amongst attributes. In [6] Mohamed Tabris *et al* proposed a hybrid behavior analysis technique for spam filtering. This technique is based on text or html in the body of the incoming email and in its header. In order to improve detection rate of approach attachments and images included in email are also considered. In [7] Nadir Omer Fadl Elssied *et al* proposed a concept using K-means clustering and Support Vector Machine in hybrid way for spam classification. This model improved detection rate and also reduced the time cost and false positive. This method gave 98.01% of classification accuracy, 0.04% false positive and 63.09 second for time cost. The result of proposed system was compared with Support Vector Machine based spam detection system. In [8] Mario Antunes *et al* proposed a text classification technique based on an immune inspired adaptive technique and non adaptive machine learning support vector machine. This method is also used by some researchers for spam filtering for getting better results in comparison of traditional techniques. In [9] Manjusha K *et al* used a BDT-MSVM based spam filtering approach. This method has benefits associated with support vector machine and Decision Tree.

In our proposed work we are working on Support Vector Machine and Artificial Immune System, hence some background information is discussed as:

- *Support Vector Machine-*

Support Vector Machine based classifiers have superior classification precision in contrast to other classification techniques. SVM gives higher performance for accuracy, time and processing speed for high dimensional data. It also prevails over the pest of dimensionality problem. SVM takes its decision by using support vectors. For creating support vectors it uses training data set [2].

Linear classifier of Support Vector Machine tries to exploit the limits of their decision borders for reducing the generalization error. Binary classification through SVM is simple and initially SVM was developed for same. Extension of SVM for multi-class classification is quite difficult then original design. Key concept behind Support Vector Machine is statistical learning theory. Statistical learning theory is used for determining location of decision borders. Decision borders are responsible for finest partition of classes. Key concept of very simple Support Vector Machine is shown in Fig 2.

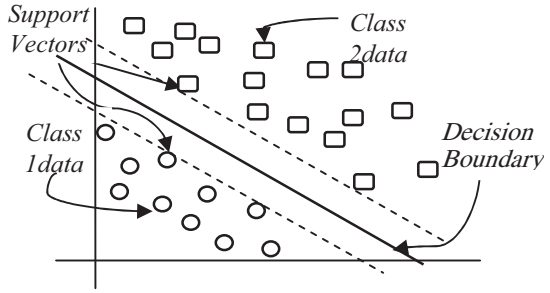

Fig 2. An overview of the SVM Classifier

In traditional two classes separator Support Vector Machine infinite numbers of decision borders are selected as support vector in order to minimize the generalization error. The closest data points are used to determine unknown class. These points are known as support vectors. It is desirable that selected decision boundary will leaves the maximum margin between the two classes. Margin between two classes is calculated as the summation of distances to the hyper plane from the nearby points of two classes.

When a problem arrive where two classes are not linearly distinguishable then Support Vector Machine tries to find the hyper planes that maximize the margin. It is also desirable from this hyper plane that the misclassification error should minimum. Settlement of margin and misclassification error is done by a constant. This constant is defined by user on the basis of various experiments conducted on the sample dataset.

- *Artificial Immune Systems-*

The Immune System evolved to become an extremely complex resistance system that has the capability to identify foreign substances (pathogens) and to differentiate between harmless and harmful [10]. Immune System is decomposed in two main layers of resistance i.e. innate and adaptive. Innate recognizes précised substances and its conduct is similar to all individuals of the same species. Adaptive layer is able to learn and to identify new forms of anomalous pathogens that regularly change during the time; hence it provides an extremely complicated adaptive form of pathogen identification. The Immune System is also supported by an intricate situate of cellular structures which is divided into small peptides by Antigen Presenting Cell (APC). Artificial Immune Systems (AIS) is an adaptive system inspired by biological immune system and it is based on theoretical immunology. This theoretical concept can be applied for solving categorization problem [11].

Oda et al. [12] proposed an immunity system for spam filtering. In his work, antibodies which offence the antigens were created by using regular expressions. Throughout this process each antibody may compare quantity of antigens (spam), that decreased the antibodies (features) set. In [13] [14] Ruan *et al.* proposed a concentration based feature construction (CFC) technique. In this technique self-concentration and non-self-concentration were determined by using terms in self and non-self-libraries. The terms within the two libraries were merely selected according to the tendencies of terms. In order to explain working of Spam Filter in actual manner four filters are used in their work along with the restrictions or demerit of them. These filtering techniques are discussed as:

A. Blacklist based filtering-

Blacklist is a list IP address which is actively used for spamming. This approach detects and provides protection against active spammers. Blacklist is used for jamming well-known spammers but there are some major drawbacks associated with this approach given as:

Demerits:

1. Manual updating is required if the filter is applied in isolation, hence user requires to update the blacklist continuously for better blocking.
2. There is a danger of blocking a legitimate email if the email sent from a blacklisted IP address.

B. White-list based filtering-

White-list is a list IP address which is actively used by authorized email senders. This approach allows only email which is sent from an authorized sender. Demerits of this approach are given as:

Demerits:

1. Trusted user and his IP address should be known for preparation of white-list in advance.
2. Continuous update is required, and it leads to manual work.

C. Content based filtering-

For spam filtering Bayesian approach is best known content based filtering technique. In content based filtering technique major focus of system is on contents rather than other information associated with incoming email. Hence accuracy rate of any content based filtering technique is better than some other methods; but due to some limitations these types of methods have some demerits discussed as:

Demerits:

1. An initial setup is required for text categorization rules.
2. Regular update of spam dictionary according to user is required.
3. Definition of spam changes according to user hence every user requires different dictionary.

D. Forging based filtering-

Recently spammers are using forged address for sending spam email. They use forged address in from field to bypass white list and blacklist based techniques. Forging based spam filtering rule verifies domain name in the 'From' field of incoming mail with domain name of actual mail sending server. Whenever it gets matched incoming mail is marked as legitimate otherwise it is sent in spam folder. There two types of forging methods are known to us. First is email based forging and second is server based forging. This method also has some demerits discussed as:

Demerits:

1. This method uses IP address based technique, contents of incoming mail are not analyzed whether it is malicious or not.

III. PROPOSED SYSTEM

A detailed discussion is carried out in previous section about existing spam filtering techniques. We find that content based filtering technique is very popular and extensively used. In our proposed system we have developed a Hybrid System by considering the advantages of Support Vector Machine and Artificial Immune System. In this system we have applied parallel filtering model where both of these strategies are working independently and their results are combined with our proposed formula. Support Vector Machine has its own advantages where Artificial Immune System provides its own benefits. When we consider result of both the techniques jointly it gives good filtering results. Filtering results are also affected by intermediate method like term selection and feature extraction. Selection of these two methods may cause a major difference between the results. In our proposed work we used Information Gain (IG) for term selection and Local Concentration (LC) based method for feature extraction.

A detailed structure of proposed system is shown in fig 3. In our proposed system five processing stages are involved to generate final results. Each of them is discussed as:

A. Preprocessing-

In this phase data is preprocessed. When setup is working with real time spam filter incoming email is processed and when working in an experimental environment sample datasets are preprocessed. In this phase by using string tokenizer a dictionary of words is created. Some irrelevant words may be discarded during this phase. After this stage processed data is passed to next stage of the system.

B. Term Selection-

For getting better results from the spam filter term selection strategy plays a major role. This stage of the proposed system selects the terms from processed data and passes it to feature extraction phase. In this work Information Gain is used as term selection method. A detailed structure of experimental setup is taken from [2] in our work.

C. Feature Extraction-

At this stage spam filtering system analyzes the selected terms for feature extraction. Features are key points of selected terms. Extracted feature from the terms play key role in spam classification. There are many methods available for feature extraction. In our proposed work we are using Local Concentration based approach. A detailed structure of experimental setup is taken from [15] in our work. This approach gives better results in comparison of others.

D. Classification-

At this stage two classifiers are used in a parallel way with aimed to get higher accuracy and shorter response time, however a system which forms serial combinations of filters may take higher time than parallel. In our system one is Support Vector Machine [5] and another is Artificial Immune System [16]. Detailed working of both of them is discussed in earlier sections of this paper.

E. Result Calculation-

The results given by SVM and AIS are stored in an array of binary numbers. Elements stored in the array (0 or 1) specify the outcome of the classifier. Zero represents spam and 1 shows that the given message is legitimate. Then weighted average is computed.

It assumed that weights are accurate. For calculating mean value (M) of both results following formula is used:

$$M = \frac{\left\{\frac{\alpha_1 * F_1 + \alpha_2 * F_2}{F_1 + F_2}\right\}}{\left\{\frac{\beta_1 * \bar{F}_1 + \beta_2 * \bar{F}_2}{\bar{F}_1 + \bar{F}_2}\right\}}$$

In above formula α is spam co-efficient, β is legitimate co-efficient, F_1 & F_2 are filter rule corresponding to SVM and AIS, and M is spam mean.

IV. EXPERIMENTAL SETUP

We conducted all the experiments on a PC with Intel(R) Core(TM) i5-4200U CPU @1.6 GHz and 4GB RAM using MATLAB.

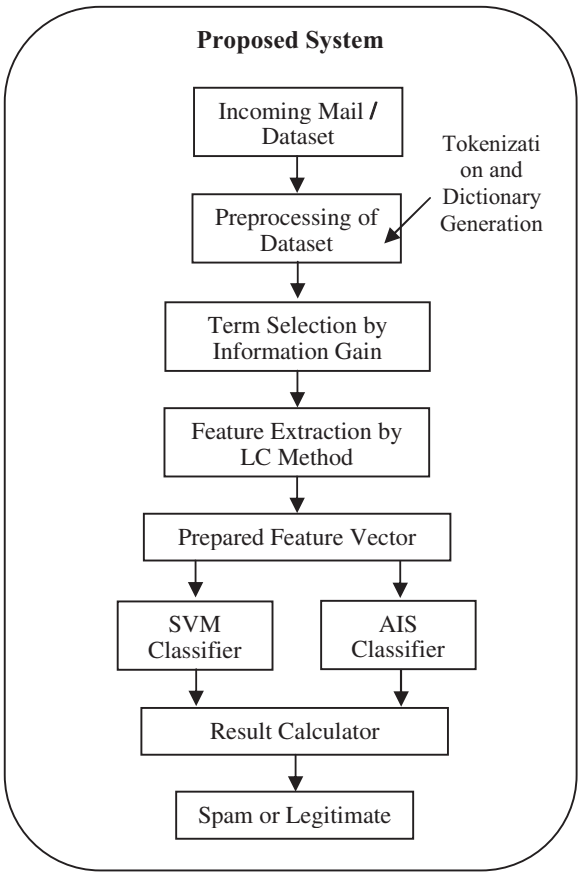

Fig 3. An overview of Proposed System

We conducted our experiments on four benchmark corpora PU1, PU2, PU3, PUA. The corpora are pre-processed with elimination of HTML tags, attachments, and header fields. In all PU corpora, the duplicates were separated because it might cause over-optimistic results in experiments. In PU1 total 1099 messages are considered out of which, 481 messages are spam and remaining 618 are legitimate. In PU2 total 721 messages are considered out of which, 142 messages are spam and 579 are legitimate. In PU3 total 4139 messages are considered out of which, 1826 messages are spam and 2313 are legitimate. In PUA total 1142 messages are considered out of which, 572 messages are spam and 570 are legitimate. All the messages are available in pre-processed form and also available in English. These messages are received by the author of [17] over 48 months and 34 months ago respectively.

V. RESULT ANALYSIS

In this section we have presented a comparison between our proposed work with Support Vector Machine and Artificial Immune System respectively. For all the methods our main focus of comparison is on accuracy. Comparisons of results are shown in Table II.

TABLE II. RESULT COMPARISION OF PROPOSED SYSTEM WITH SVM & AIS CLASSIFIERS

Method → / Dataset ↓	SVM		AIS		Proposed System	
	Spam	Legit	Spam	Legit	Spam	Legit
PU1 (Total 1099 message)	434	665	428	671	453	646
PU2 (Total 721 message)	149	572	156	565	141	580
PU3 (Total 4139 message)	1828	2311	1809	2330	1804	2335
PUA (Total 1142 message)	570	572	572	570	569	573

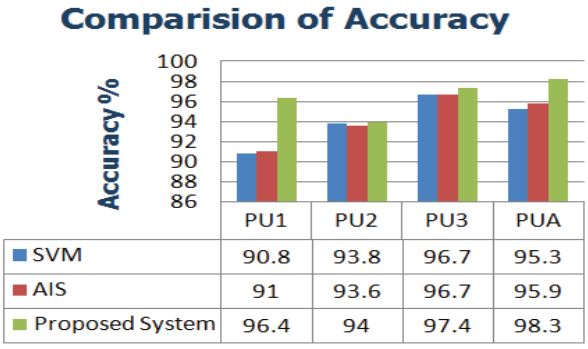

Fig 4. Comparison of Accuracy

VI. CONCLUSION

In this paper we have analyzed various existing spam filtering approach with their demerits. After analyzing them we proposed a system which has properties of Support Vector Machine and Artificial Immune System both. We have conducted all the experiment with existing data sets in a controlled environment. Future extension of this work may include a real time application of our proposed work for effective spam filtering.

REFFERENCES

[1] Ali Ahmed A.Abdelrahim, Ammar Ahmed E. Elhadi, 3Hamza Ibrahim, Naser Elmisbah "Feature Selection and Similarity Coefficient Based Method for Email Spam Filtering", Inter national Conference on Computing, Electrical and Electronic Engineering (ICCEEE) 2013.

[2] Kunal Jain, Amrit Pal, Manish Sharma, Sanjay Agrawal, "Impact of Spam on the Environment & its Prevention", IEEE Inter national Conference on Convergys of Technology, Pune, 2014.

[3] Six Different Types of Spam and How to Avoid Them http://www.allspammedup.com/2010/09/6-different-types-of-spam-and-how-to-avoid-them [Online].

[4] Email Statistics Report, 2013-2017 (Radicati Group) http://www.radicati.com/wp/wp-content/uploads/2013/04/Email-Statistics-Report-2013-2017-Executive-Summary.pdf [Online].

[5] Muthukaruppan Annamalai, Ankur Jain, Vaishnavi Sannidhanam, "A Novel Hybrid Approach to Machine Learning", Study Documents, Department of Computer Science and Engineering, University of Washington.

[6] Mohamed Tabris, Youssif B, Alnashif, and Salim Hariri, "Filtering Spam by Hybrid Approach", Orlando, United States, October 15-18, COLLABORATE.COM 2011.

[7] Nadir Omer Fadl Elssied, Otman Ibrahim, Waheeb Abu-Ulbeh, "An Improved of Spam E-mail Classification Mechanism Using K-Means Clustering", Journal of Theoretical and Applied Information Technology, Vol. 60 No.3, February 2014.

[8] Mario Antunes, Catarina Silva, Bernardete Ribeiro, and Manuel Correia, "On Using An Ensemble Approach of AIS and SVM for Text Classification", WACI 2010.

[9] Manjusha K, Rahul B. and Shahabas S, "An Efficient Method of Spam Classification by Multiclass Support Vector Machine Classifier", pp. 102–110, ICDMW 2013.

[10] K. Murphy, K. Murphy, P. Travers, M. Walport, and C. Janeway, Janeway's immunobiology. Garland Pub, 2008.

[11] L. de Castro and J. Timmis, Artificial Immune Systems: A New Computational Intelligence Approach. Springer, 2002.

[12] Oda, T., & White T., "Developing An Immunity to Spam. Lecture Notes in Computer Science", 2723, 231–242, 2003.

[13] Y. Tan, C. Deng, and G. Ruan, "Concentration Based Feature Construction Approach for Spam Detection," in Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN2009), Atlanta, GA, pp.3088–3093, Jun. 14–19, 2009.

[14] G. Ruan and Y. Tan, "A Three Layer Back-Propagation Neural Network for Spam Detection Using Artificial Immune Concentration," Soft Computing , vol. 14, pp. 139–150, 2010.

[15] Wanli Ma, Dat Tran, and Dharmendra Sharma, "On Extendable Software Architecture for Spam Email Filtering", IAENG International Journal of Computer Science, vol. 4, pp. 129–136, 2013.

[16] Mayank Kalbhor, Shailendra Shrivastava, Babita Ujjainiya "An Artificial Immune System with Local Feature Selection classifier for Spam Filtering" IEEE-31661 4th ICCCNT, Tiruchengode, India 2013.

[17] Androutsopoulos, I., Paliouras, G., &Michelakis, E. "Learning to Filter Unsolicited Commercial e-mail", NCSR, Demokritos Tech. rep. 2004.