THE PROCESS OF MULTILEVEL FILTERING ON SPAM MAIL

B. BALA KRISHNUDU* & DR. V. RAGHUNATHA REDDY

Departmen. of Computer Science and Technology, Sri Krishna Devaraya University, Anantapuramu, Andhra Pradesh, India

ABSTRACT

The E-mails on spam have been categorized on the basis of the content and the source address that are receiving. The Classification or Categorization has been performed on the basis of abstraction of contents. The Classification of E-mails gets puzzled with the issues of reduction with the dimensionality, time complexity, categorization exactness (or) classification accuracy. The Content Classification has fairly been based upon the multimedia information and text categorizations on textual information get presented on the E-mail messages. The categorization or classification of textual and multimedia information on E-mails has been done in various levels like normal, informatics, menacing to acquire higher accuracy. The classification or categorization of text that has become the vital technique for managing and handling the text data. The main aim of the categorization or classification of text is to organize the document into various predefined categories. Every document on multirole precisely one or without any a definite category or group

KEYWORDS: Email, Spam, Multi-Level Spam Filter, Multi-Objective Spam Filtering, E-mail , Classification

Received: Jan 09, 2021; **Accepted:** Jan 29, 2021; **Published:** Feb 23, 2021; **Paper Id.:** IJCEITRJUN20212

INTRODUCTION

Mr. Lahmadi and Festor have jointly been proposed filter on the social network to be performed the trivial message services-based spam thru the filters which are Bloom Filters and content hashing Filters [1]. This Joint-venture Filtering plan also depends on the centralized server to construct the social networking amongst end-users.

Mr. Hameed.et.al. (2011) have been gone through the social network on who receives E-mails and diminish the spam outside the social-circle that also gets alleviated or lessened the consumption of bandwidth by the spam. To get resisted spam, Malwares and Phishing[2] three URL's, the pen-men or authors, Thomas.et.al.2011, have been implemented a system which is real-time one being incorporated the aggregation of URL's, the feature collection, the Feature Extractions, categorization or classification or segregation or division.

Li and Shen on 2011, have been implemented a frame-work for social networking-based spam-filtering. It detects the junk E-mails [12] with the concern of End-user social features and networks therefore the regulator and junk E-mails get segregated. On social networking the trustiness exploits jointly in the name of filtering. The Spam's Reporters trust worthiness will be used to gather the righted or rectified reports of spam and get detected the Sybil assails as well [3].

Shakar and Karypis on 2000; Mr. Yang and Pedersen on 1997, have been implemented an algorithm to mitigate the salient feature space sans making sacrificed eccentric (remarkable) classification accuracy. The efficiency of the system fairly relied on the training dataset quality.

Aery and Charavarthy on 2005, have suggested a mining approachment on the basis of graph theory for the classification of E-mails. The Structure or Architecture or Construction of framework [1] gets provided such “ism” got retrieved from the folder of Pre-Classified E-mail and the same kind of thing can efficiently be laboured or employed for categorizing the incoming E-Mail Messages.

The fundamental approachment of content-based Spam’s filtering will be the list of static keywords and still it gets made easier for spammer to avoid filtering by making a small or petty adjustment over such message. The content-based approachment is the second category that embodies ML(machine learning) based approachment like Bayes filters, Decision trees, SVM’s, Bayes classifiers[9] and blending of such techniques[8]. This kind of approachment on which a learning algorithm has been employed to search the attributes or characters or qualities of the spam and of licit E-mails. The Future Messages can automatically or usually be classified as to be spam and to be the righteous – E-mails or somewhere among by then. The next category of content based approachment will be associative based Spam Filtering. Suppose the spam-E-mail detects by the End-user where as other End-users in the campaign or camp shuns the spam later on by querying others to notice it as if such received E-mails are spam or not spam [10].

RELATED WORK

The current and present world has utterly been impacting by the technology of internet and the folks have been infected and influenced and influenced on amount of large scale by using the services of internet on their regular activities and business activities. If you use it for positive deeds it can give you a pleasant consequence. Whereas if you use it for negative deeds, it gives you a way to the hellish. What is happening around the area of internet where have been several spam e-mails got sent by the different end users from various places or through some of the programs are automated. Actually, on the functionalities of the business, the bearing and carriage of spam e-mails or accompanying of spam e-mails get dismissed the efficacy of business peoples who are being operated or worked on the e-mails of business kind. Normally, the people classify spam-emails on the basis of mails content or in some cases it depending on the e-mail categorization or classification based on the perspective of the user or customer who sends the e-mails frequently or regularly.

The dimensionality that is one of the prominent problems in organizing or classifying thee-mails. The classification or categorization will be done in two ways.

- The Taxonomy usage
- The Feature Size.

The feature deduction or reduction may get attained from different properties or components and classes label that got belonged to the Domain Ontology. To make reduced the dimensionality, one must be concerned the classes of the feature or Domain Ontology. The time complexity has been alleviated for the classification approach by the feature of reducing the dimensionality.

Domain Ontology

Ontology can represent as the philosophical study of Conceptualization. An Ontology being represented in terms of taxonomic hierarchies of classes and class definitions and principles and nexus among the classes. To make specify or identify the conceptualization, one who needs or requires to represent the axioms or doctrines for the terms that defined.

The Ontology plays a vital role in making captured the domain knowledge and make provided a regular and basic common understanding. Actually, Ontologies consist of taxonomy, the class hierarchy, the base of domain knowledge and the nexus between the classes and instances. Ontology would have various relations that rely upon the schema or builder of taxonomy, and it would have various restraints that relying on the builder of Ontology. Ontologies allows for the data semantics that can be understood by the machines and make facilities the search, exchanging, the knowledge integration for B2B and B2C at E-Commerce. By employing or using the semantic data, the usage of Electronic- Technology will be easier and smoothed.

Ontologies quests more effectively and efficiently rather than the methods or patterns which are traditionalized. Naturally, Ontology exists in the hierarchical description of vital ideas or concepts in the domain and the elucidations or depictions of concern properties of each and every idea or scheme or concept. Conventionally, Ontologies have been built or developed by the technocrats who trained highly or greatly and the domain specialists who don't have such knowledge regarding to the computer software. The construction of Ontology has been a time – consuming part and whopping task. Ontology tools also need customers or users to get trained in representing of knowledge and logical sensing and technicality.

The suggested domain Ontology existed in the following classes/divisions such as

- Info or Informative
- Generic or General
- Threatening or Attacking or Menacing or assailing
- Severe or Extreme

The suggested approachment will have different levels or stages or states of spam filter named as

- Pre-processing or initiate the process at early stage
- The Identification of Multi-Level Feature
- Multi Objective of Spam Filter

1. THE ARCHITECTURE OF THE SYSTEM FOR THE FILTERING OF MULTI LEVEL SPAM ON EMAIL

The approachment would have different levels of filtering of the spam [15], represented as above. Each and every functional based component that has been elucidated in the below picture or Figure 1

Figure 1: Filtering of Multi Level Spam

1.1 Pre-processing

On the Pre-processing level or state, initially different features of textuality have been recognized on the mail. That refers, the subject would represent on the mail. At Second Level, a bunch of all attachments and their types have been recognized. The features and the classes have been recognized for each and every attachment. The bunch of features has been recognized gets employed at the next level to perform the feature extraction.

Algorithm for Pre-Processing:

Input : Email represents as 'Em'

Output : Feature Set represented as 'Fs'

Step 1 : Get Started

Step 2 : The Extraction of Subject Feature represented as

$Sf = \sum \text{the terms belong to Email Subject}$

Step 3 : The Feature Extractions shown on the container of the mail.

$\text{Mail Container(MC)} = \sum \text{Terms belong to the container of the Email}$

Step 4 : Get extracted the attachment (At) from mail and belongs to Email.

$At = \sum \text{Attachments belongs to Em}$

Step 5 : Feature Set **F_s**= set of features represented as **SF, MC, At**

Step 6 : Get Stopped

The Pre-Processing stage or state has been elucidated with an instance like

A. End use X sends an email to the user Y

B. It would get classified on the receiving mails like

- The Subject of the concern E-mail.
- The Container of concern E-Mail.
- The Attachments of concern E-Mail.

In the suggested or proposed model, the words of spam have been pre-defined and they have been listed on the Table 1 in below.

Table 1: The terms of spam got engaged in the suggested framework

For You	Get Accepted Credit Card	Get Access Now	Extra Income	Making Amazed
Got Applied On-Line	Auto Email Discarding	As Saw Upon	Auto Email Discarding	Get Amazed
As A Member	Got A Bug on A/C of Big Amount	The Address of a Billing	Emails on Bulk	Got Bought Directly
Get Called Free	Get Called Now	Make A Cancelled at Anytime	The Bonus on Cash	Get Checked (Or) Money-Order
Dear Someone	Make It Today	For Get Freed	Pay	Discount

In the suggested or proposed framework 150 terms of spam have been corresponded to the account. Whenever the recipient Y gets received the mail, the mail-contents got segregated or separated into terms and get cross- validated with terms of spam on Table 1. Still, it also assays or estimates the attachments with regard to Email and categorized the attachments into classes, like

- Has been _ Text
- Has been _Image
- Has been _ Audio

1.2 The Identification of Multi-Level Feature

The Identification of Multi-Level Features has been the process of making identified the features at various stages of Domain Ontology. The method manages a bunch of classes, attributes or characters or qualities or properties and values for various stages of spam [5]. For each and every class of spam stage, there have been several classes and the properties got handled or organized that have been laboured to be identified the features of various stages that held on the processing of mail. For the message or content that related to multimedia, the method maintains or handles different kinds of set of class-features and computes or calculates the similarity metrics based on the values of feature descriptor.

Algorithm for Multi-Level Feature Identification:

Input: Domain Ontology Do, Multimedia Feature Set MFS, feature, Vector fv.

Output: Multi-Level Feature set MLFS.

Step 1 : Get Started

Step 2 : For each level L from the Level set (Ls) $L \in Ls$

Make Identified no. of classes and the attributes got suited from Domain Ontology

size(class)

$NCMM = \int \Sigma \text{prop}(C) \in Fv$

$I=1$

here Prop(C) is property of Class C

MLFS = {NCMM}.

End

Step 3 : For each attachment Ai from Fv ($Ai \in Fv$)

For each level Li from Ls ($Li \in Ls$)

Calculated Feature descriptor $Fd = \int \Sigma$

Calculated Attachment distance (Ad)

size(MFS)

$Ad = \int \Sigma \text{Dist}(Mfs(i), fd)$

$I=1$

MLFS = MLFS + {Ad}.

End.

Step 4 : Get stopped

Now we considered that an Email Y-get received, it's fairly a text based-Email. According to such consideration, the following taxonomy has been listed out as

- The Mail which is a Normal one
- The subject regarding to the Mail (that is Mining of the Data, Cryptography etc.,)
- The message with warnings

The third taxonomy has been sub-classified as

- The messages which are unauthorized meant by got mails from anonymous people
- Menacing mails or Mils of Threatening (classified by using words such as killing, running, eradicating)
- Server-Oriented (making targeted or aimed at concern organization)

1.3 The filtering of Multi-Objective Spam

It has been accomplished in calculating the set of multi-level features and the set of Multimedia features [4] and the availability of Domain Ontology. Various kinds of similarity metrics or measurements or calculations at various feature levels that have been calculated as

- Info or Informative
- Generic or General
- Threatening or Attacking or Menacing or assailing
- Severe or Extreme

That have been elucidated and specified on the table 2 in below

Table 2: Evaluation Metrics

Parameter	Value
No. of semantic classes	4
Classes Name	(i) Info or Informative
	(ii) Generic or General
	(iii) Threatening or Attacking or Menacing
	(iv) Severe or Extreme
No. Of user mails considered	150

For each and every stage, multiple objects of spam similarity metric has been computed or calculated by employing the Feature Vector (FV) is building and the Domain Ontology will be available. On the ground of class that has a greater MOSSM value then mail gets allocated a class and gets classified or categorized [4].

Algorithm for Multi-Objective Spam Filtering

Input: Domain Ontology Do, Multimedia Level Feature Set MLFS

Output: Class Ci.

Step 1 : get Started

Step 2 : For each class Ci from CS Ci ∈ CS

calculate Multi Objective Spam Similarity Measure (**MOSSM**)

size(ls)

$$\text{MOSSM} = \int \sum_{i=1} \text{Class.Prop} \in \text{MLFS} / \text{size}(\text{prop.class})$$

i=1

For each Multimedia Class Mi

MMS=Multi Media Similarity

size(MMC)

$$\text{MMS} = \int \sum_{i=1} \text{Dist} < \text{MMTH} / \text{size}(\text{MMC})$$

End

MOSSM = **MOSSM** * **MMS**

End.

here Prop(C) is the Property of Class C

Step 3 : Get Chosen MOSSM valued class at maximum

Step 4 : Get Stopped

2 THE PERFORMANCE EVALUATION

The Emails have been extracted or retrieved and have been stored as a dataset. The dataset has been categorized or classified by using Multiclass-Classifer [13]. The outputs have been assessed by using WEKA-3.7 tool which belongs Data mining. The attributes or Characters have been engaged or involved in the semantic classes that are classified fairly on the basis of keywords and they got listed on the table 3 in below.

Table 3: the Attributes on the classes which are semantic based

Attributes	Semantic Classes			
	Generic	Informative	Threatening	Severe
No. Of Instances	52	62	71	71
No. Of Fields	15	12	12	15

The spam classification performance [6] has been calculated or computed namely as accuracy, the ratio of false classification, the time complexity these have been the effective and active based on the performance parameters. The performance has been blended of two processes such as training and the process of testing under the processes of training, classification has not been matched to classification of testing. The below table to 4 shows how the False Positive, False Negative have been computed.

Table 4: the computation of FP & FN

	Non-Spam	Spam
Non-Spam	TRUE Negative (TN)	FALSE Positive (FP)
Spam	FALSE Negative (FN)	TRUE Positive (TP)

i) **Accuracy:** - It calculates the Spam-Weighted percentage and the Non-Spam artefacts (parts) on E-mail and It has correctly been classified or categorized.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} * 100 \Rightarrow 2.1$$

ii) **The Ratio of False Classification:**

The FP-Rate will be computed as the ratio among the “No. of Events which are Negative” that are classified wrongly and projected them positively (like False Positive) and the Total no. of Negative events [14] which are negative based (not regarding with classification).

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{FN}} * 100 \Rightarrow 2.2$$

The below Figure 2 Refers the classification for the classes which are semantic – based named as “General, Informative, threatening and Severe”.

Figure 2: Semantic Classes Classification Percentage

The below Figure 3 Refers the classification for the classes which are semantic – based named as “General, Informative, threatening and Severe”.

Figure 3: Semantic Classes Classification based on Time

The performance at whole of suggested Multi Objects of Spam Filter in referred on the table 5 in below.

Table 5: MOSF resultants at Whole

Class	Generic	Informative	Threatening	Severe
Total Instances	52	62	71	71
Correct Classification	50	61	70	69
In Correct Classification	2	1	1	2
Classification Accuracy	96.15	98.39	98.59	97.18
False classification Accuracy	3.85	1.61	1.41	2.82
Time Taken to Build Model(Seconds)	0.03	0.02	0.05	0.02

3. THE ANALYSIS OF COMPARISONS

This section refers the comparisons of results among the various kinds of E-mail filters named as category-1, category- 2, category-3 and the suggested one MOSF(Multi Objects of Spam Filter) represented as category -4[7].

Category-1: Objective

The methods that are utterly objective-type getting trouble from the FP and FN classification. The methods that are objective-type are basically relied on the filtering of content and are time- consumed, resource-challenging. They have not been accurate and acute and require continuous update to cope up or managing with the newly constructed or invented tricks of spammer. Simultaneously at other hand an existing subjective proposal possessed some disadvantages whereby the menaces or assaults or attacks from the vicious or bitchy users that make them untrustworthy & private [21].

Category-2: Based upon URL (Uniform resource Locator)

In this, the URL division corporates the raw data with regarding to the page content and the behaviour of the page. The

gathered data which is raw one that converts into Boolean values such as 0 or 1 (yes or no) on the ground of information that has been represented on the web-pages[20].

Category-3: SPOT

The 'SPOT' has been designed by employing the statistical tool that is called Sequential Ratio Test (SRT). SPOT has been a light weight compromised machine detection scheme by questing the Economic based incentives. For the attackers or assailers for recruiting the huge No. of machines that compromised. Let X_i for $i=1, 2, \dots, M$, that refers the successive observations of a random variable X is in relating to the message that are sequenced and originated from machine 'M' that is inside of the network.

Let $X_i=1$ Suppose the message 'i' from the machine has been a spam and $X_i=0$ otherwise. The Detection system assessing that the compromised machine's behaviour has utterly been different from the normal machine on the basis of messages which they sent or passed. Mainly the compromised machine creating a spam message with a greater probability rather than a normal machine.

The probability has been computed as specifically in below

$$P(X_i=1/H_1) > P(X_i=1/H_0) \quad (3.3)$$

Where H_1 – The Machine which is compromised one.

H_0 – The machine which is normal one.

It also embodies the content-based spam filter to make identified whether the 'Message' is going out on SPAM (or) ham [11]. The CT (Count Threshold) & PT (Percentage Threshold) got computed by then. The CT module has been counting the no. of messages at spam sent by the compromised system on the network. The CT computes the machine sending messages in a specified/stipulated time interval [19].

Suppose the machine count increases then it automatically decides that system has been as a system of spam. The PT gets calculated the system being sent message and being counted the no. of message of spam by the system that's compromised one and eventually the spam message percentage will be computed. The feature embodies in this category 3 (SPOT), have been

- The Ranking of Mail (it covers spammer & actress of non-spammer)
- The Rate of Participation

Figure 4 elucidates the suggested the multiple –objectives of classifier labouring domain ontology that receives 95.15% avg accuracy value of classes that are semantic based & the objective 83.14%, URL got 75.72%, SPOT gets 89.7% [18]. So, here the suggested work has been attained the greater classification exactness or accuracy in comparing with a specific category.

Figure 4: Comparison of Classification Accuracy

Figure-5: Expounds the suggested the classifier of multiple. Objectives by employing domain ontology obtain the petty based classification with falsified and ratio of 4.0% as compared to the other techniques of classification.

Figure 5: Comparison of False Ratio

4 SUMMARY

The Filtering of Multi - level spam has been suggested with the classifier of multiple objectives with the alleviated or diminished dimensionality [16]. The method being performed the pre-processing that retrieved various levels of features which belonged to the text content and multimedia content. At the next level or stage the mails have been segregated into the classes that are semantic based like General, Info, Attacking or Threatening or Assailing and Severe or Extreme [17]. The Classification performs upon the classes that are always being semantic. The Consequences referred that the suggested method generates or produces the effective and ability-based classification with the greater accuracy or exactness & lessen the falsified ratio.

REFERENCES

1. Jianying Zhou, Wee-Yung Chin, Rodrigo Roman, and Javier Lopez, (2007) "An Effective Multi- Layered Defense Framework against Spam", *Information Security Technical Report* 01/2007.
2. Rafiqul Islam, Jemal Abawajy, "A multi-tier phishing detection and filtering approach (2013)", *Journal of Network and Computer Applications*, Volume 36, Issue 1, January, pp. 324–335.
3. Xiao Mang Li, Ung Mo Kim, (2012) "A hierarchical framework for content-based image spam
4. Filtering", *8th International Conference on Information Science and Digital Content Technology*
5. (ICIDT), Jeju, June, pp. 149-155.
6. Z. Wang, W. Josephson, Q. Lv, M. Charikar and K. Li. (2007) *Filtering Image Spam with near- Duplicate Detection*, in *Proceedings of the 4th Conference on Email and Anti-Spam CEAS*.
7. Duncan Cook, Jacky Hartnett, Kevin Manderson and Joel Scanlan (2006), "Catching spam before it arrives: domain specific dynamic blacklists", *ACSW Frontiers '06 Proceedings of the 2006 Australasian workshops on Grid computing and e-research - Volume 54, January*, pp. 193-202.
8. Anirudh Ramachandran, Nick Feamster, and Santosh Vempala, (2007) "Filtering spam with behavioral blacklisting", *Proceedings of the 14th ACM conference on Computer and communications security CCS'07, NY, USA*, pp. 342-351.
9. Damiani E., Vimercati S. D. C. d. et al., (2004) "An Open Digest-based Technique for Spam
10. Detection", *San Francisco, CA, USA*, pp. 1-6.
11. Harris Drucker, Donghui Wu, and Vladimir N. Vapnik, (1999) "Support Vector Machines for Spam Categorization", *IEEE Transactions On Neural Networks*, Vol. 10, No. 5, September.
12. Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz (1998), "A Bayesian Approach to Filtering Junk E-Mail", *Learning for Text Categorization: Papers from the 1998 Workshop, Madison, Wisconsin, AAAI Technical Report WS-98-05*.
13. Dat Tran, Wanli Ma, Dharmendra Sharma, and Thien Nguyen, (2007) "Possibility Theory-Based Approach to Spam Email Detection", *IEEE International Conference on Granular Computing*.
14. Clotilde Lopes, Paulo Cortez, Pedro Sousa, Miguel Rocha, and Miguel Rio, (2011) "Symbiotic filtering for spam email detection", *Expert Systems with Applications: An International Journal*, Volume 38 Issue 8, August, pp. 9365-9372.
15. Oda, T. and T. White. (2005) *Immunity from Spam: An Analysis of an Artificial Immune System for Junk Email Detection*. in *4th International Conference on Artificial Immune Systems (ICARIS)*.
16. M. Uemura and T. Tabata, (2008) "Design and Evaluation of a Bayesian-filter based Image Spam Filtering Technique", in *Proceedings of the International Conference on Information Security and Assurance (ISA)*.
17. Dat Tran, Wanli Ma, and Dharmendra Sharma, (2009) "A Novel Spam Email Detection System Based on Negative Selection", In *Proceedings of the 4th International Conference on Computer Sciences and Convergence Information Technology (ICCIT'09)*. Los Alamitos, CA, 2009, pp. 987-992.
18. Elena Zheleva, Aleksander Kolcz and Lise Getoor, (2008) "Trusting spam reporters: A reporter-based reputation system for email filtering", *ACM Transactions on Information Systems (TOIS)*, Volume 27 Issue 1, December.
19. P.K Panigrahi, (2012) "A Comparative Study of Supervised Machine Learning Techniques for Spam E-mail Filtering", *Fourth International Conference on Computational Intelligence and*

20. *Communication Networks (CICN)*, Mathura, Nov, pp.506-512.2
21. http://www.securelist.com/en/analysis/204792321/Spam_in_November_2013
22. <http://www.sans.org/reading-room/whitepapers/email/is-affect-us-deal-spam-1111>
23. <http://www.spamlaws.com/spam-stats.html>
24. <http://www.spamexperts.com/en/news/motivation-spammers>
25. <http://www.spamhelp.org/software/>
26. V. Siva Sankaran & N. Kannan , “Increasing Brand Reputation through Viral Marketing “, *International Journal of Sales & Marketing Management Research and Development (IJSMMRD)*, Vol. 6, Issue 1, pp. 1-8
27. Ulka M. Bansode & Gauri R. Rao , “Study of Various Anti-Phishing Approaches and Introducing an Improved Method for Detecting Phishing Websites “, *International Journal of Computer Science and Engineering (IJCSE)*, Vol. 2, Issue 4, pp. 151-156
28. Prashant K. Shah, Akshar L. Muley & Komal J. Dalwadi , “The Study on Use of Information and Communication Technologies to Support Deliberate, Repeated and Hostile Behavior by an Individual or Group that is Intended to Harm Others” , *IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET)*, Vol. 2, Issue 5, pp. 15-24
29. Ahmed Saleem Abbas , “Side Effects of Computer Anti-Viruses “, *International Journal of Computer Science and Engineering (IJCSE)*, Vol. 6, Issue 4, pp.11 - 20