Research Article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Employing the Word-Net Taxonomy for Tree - Structure Spam-Classifier for Filtering of E-Mail Spam

B. Bala Krishnudu* and Dr.V. Raghunatha Reddy

Dept. of Computer Science and Technology, Sri Krishna Devaraya University, Anantapuramu, Andhra Pradesh - 515001, India.

ABSTRACT

The Current World being Penetrated to numerous communication devices and the methods to get sent and received the data. By enormously, the internet and the electronic devices which are communication based, being utilized. The communication which is electronic based, the E-mails have commonly utilized the communication protocol at most across the globe. The frameworks of spam Email at large amounts, are used the keywords to get recognized the SPAM msgs. These kinds of keywords should compose as incorrect spellings. For an instance, “tutor” for “tuter”. For time to time, the incorrect spellings have been changed, therefore, through the spam detection system, the righteous update must occur. It is very tough task to make anticipated that the total feasible incorrect spellings for the yielded keywords get embedded to the Blacklist. In this chapter, the feature approximation of textuality and detecting the object these two methods have been suggested for making progressed the classification of Emails.

Keywords: WordNet, Tree structure classifier, Spam; E-Mail

***Corresponding author**

B. Bala Krishnudu

Research Scholar,

Dept. of Computer Science and Technology,

Sri Krishna Devaraya University, Anantapuramu, Andhra Pradesh - 515001, India.

Email - balakrishnudu81@gmail.com

1. INTRODUCTION

The E-mails may have the msgs based conversational at simple and incorporate the attachments at large scale. The feasibility of making passed the data three the communication on-mail is so easy rather than the voice-over communication and this kind of scenario are regularly being used by the extremists in various/different ways.

Different types of msgs which are untrusted being sent by the intruders thru the communication on mail. The data may encrypt or in the form of coding and decoding modes (or) other typical language modes as well. The intellectual based approachments are needed to crack such malicious mails in such manner. Not as only text data, the symbolic-data also be feasible to represent in the form of images. Some of the end-users which are unauthorized (or) anonymous may capture/retrieve the meaning of concern images/pictures and the communication amongst themselves. Through the attachments, the mails which are corrupted/malicious/stunk, must identify and filter from making delivered to the people (or) to make found concern malicious crooks.

2. THE WORDNET OVERVIEW

For the English language, the word-net has been served as the lexical- database. The English words compile into the bunch of synonymous, have been called as the “synset”. They would register the numerous relations amongst the bunch of synonyms or their members by making provided a short-definition and the examples/instances for exercising / use / operation / usage. The word-net represents as the blending of dictionary and thesaurus. The main object of this tool is getting analyzed the automatic – text, making applied in the area of AI. Any End-users can down-load the DB and Software tools with licensed versions by the BSD-style from the website. The End-users get accessed the files belongs to lexicographer and the compiler represented as Grind, on free basis for making generated the Distributed Databases.

The word-net embodies the categories, based on lexical, named as, nouns, verbs, adverbs and adjectives, but it avoids the conjunctions, determiners, prepositions and test of other function based words.

In simply, the word-net has been the English-words Database. The Database of English words link/ connect together by its semantic nexus. At eventually it is a kind of super power based dictionary/thesaurus with the structure of graph based.

The 'synset' have been the words from the same category/class which is a lexical one that have roughly been synonymous and categorized into synsets. The synsets contain simplex words and collocations that represent as "stand out" and "put up" are the examples of it. The numerous forms of polysemy word's have been allocated to various "synsets" .The synonymous have been the words that contains similar meanings. A synonym set (or) a bunch of synonym (or) Synset has been the passel of synonyms. So, a Synsets get corresponded to the concept based abstract.

Now, we reckon the term/word 'sport', for this 'specified term', the WordNet would give seven bunch of words/terms that shown/visualized in the following Figure1

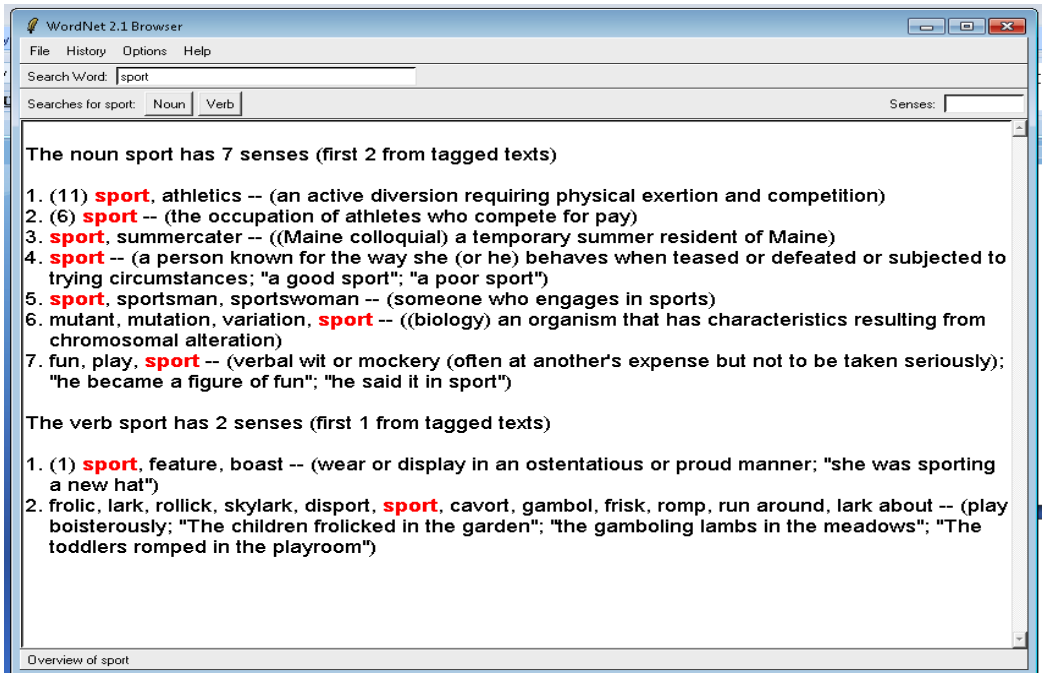


Figure 1: The instances for Synset employing the WordNet taxonomy

For making classified the mail efficiently, the WordNet taxonomy is get employed in the suggested framework.

3. THE SUGGESTED METHODOLOGY/PROCESS

The Configuration/framework/structure/construction for the suggested tree structure based spam classifier that has been mentioned at Figure 2. The steps engaged in the process of suggested, have been referred in below[1]

Step-1 : Getting Pre-processed

Step-2 : The textual feature approximation utterly recursive based.

Step-3 : Detecting the object & making mapped

Step-4 : Employing threshold technique for making the classification

The approachment is extracted the numerous features, named as, textual-based and the multimedia –based from the E-mails as a i/p. The method of multi feature approximation performs o the features which retrieved while the spam weight also calculated. On the basis of spam-weight, the method is classified the E-mail into the spam (or) Genuine E-mail. The Whole-process/total method includes numerous stages. They have been mentioned at top meant by various steps. The mentioned structure/framework represented clearly in below, expounds the model of intelligent feature approximation for making filtered the SPAM[2].

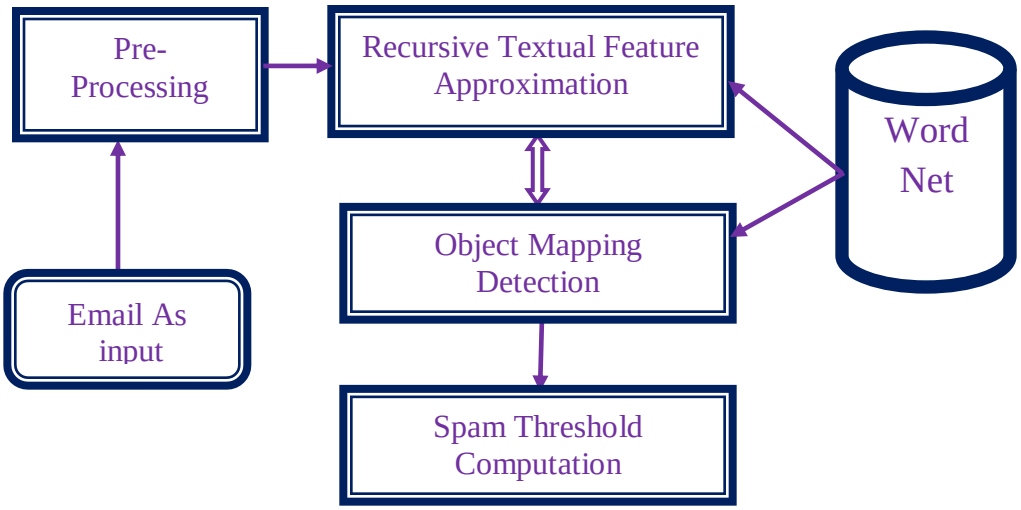


Figure 2 spam classifier using word net taxonomy

3.1 Pre-processing

It is the process of getting obtained and extracted the content of textuality and the content, those have been attached to the mail. At starting/inception, this process gets extracted the content of textuality from the E-mail where as it recognizes. The multimedia attachments catalogue (list) and gets retrieved them for making processed the mapping of objects. The features whatever represented on the Email will get retrieved to make performed the filtering of the spam.

The Algorithm based on pre-processing will handle the E-mails from the numerous resources as i/p and the results mentioned as o/p in terms of textutal and multimedia attachments. We now named the o/p as Text Set. The Text-Set as been the bunch of msgs based on texts from the i/p- E-mails. The Multimedia Set represented as the bunch of multimedia msgs from the i/p as well. As a initial-step, the content of the textuality from the mail-container and the subject from the mail, have been retrieved and appended to the Text-Set. Simultaneously, the content of multimedia based, in the attachments have been appended to the Multimedia Set [3].

Suppose, the mail could have numerous attachments, it requires to extract the attachment – content on the basis of its kind, it should append to the Text-Set or Multimedia-Set. So, the Extraction in loop for every attachment and the whole /total attachment content segregates and adds to their concern sets.

Input: E-mail (The subject on the mail, the content from the container of the mails, the attachments)

Output: 1) Text-Set , that embodies all the msgs based on the text & the line of subject on mail and its attachment as well.

2) Multimedia Set that includes the content of whole multimedia in the container and the attachments.

3.1.1 The Algorithm of Pre-Processing

Input : The E-mail

Process/method : Gets extracted the content of textuality & content of multimedia attachment

Output : Text set, multimedia set.

1. Get Started
2. Get Extracted the subject from the E-mail
3. Get Extracted the content of textuality based represent on the container of mails.
4. Append the retrieved subject & content to Text-set.
5. Makin g found the numerous attachments on the mail.
6. For Every attachment must follow

   6.1 - if/suppose the attachment contains an image, while

   6.1.1 Make it added/appended to the multimedia-set

   Else

   6.2– if the attachment will be a document while

   6.2.1 Get retrieved the text & make it appended to the Text-set
7. Get saved the multimedia set & Tet-set
8. End.

Now, we see the sample of E-mail in below

Figure 3 the Email on sample

The E-mail on sample represented at top 3 figure, the resultants as Text-Set and /or Multimedia-Set have been projected in below.

Text Set = {Hi, Good, afternoon, This, is, the, test, mail, Welcome, all, message}

Multimedia Set = {sample.img}

3.*2 Textual feature approximation based on recursive*

It has been the process/method of getting recognized the relative terms that belongs to E-mail Text and its meanings that are symbolic-based from the word –dictionary. The Words (or) terms (or) Synsets recognized from the word net taxonomy have been appended to the Text-Set, retrieved from the E-mail and it performs in a recursive way to have the symbolic terms/words that are similar based[4].

Input **:** The Text-set obtains from the pre-processing Algorithm which is previous one.

Tools Used **:** Word net – Dictionary

Output **:** Suited - Text-Set –the bunch of Synset – words the Word-Net dictionaries which have Synonimally identical/replicable to the text as i/p in the Text-Set.

The Text-Set uses/employs as i/p in the algorithm of recursive-text. It separates the E-mails as Synsets by making compared the Text-Set with the dictionary and calculates the weight of spam (computed employing spam-weight Algorithm).

3.2.1 The Algorithm of Recursive Text

Input : The text-set, Word net dictionary

Process/Method : Recognize the Sysnets from the text set and calculates the weight of spam.

Output : The Weight of spam, suited text-set

1. Get started
2. Reads Text-set obtains from the pre-processing
3. Every object/item in the Text-set accomplishes the following steps.
   - 3.1: Recognize the total sysnet employing the word-net dictionary
     - 3.1.1: Accomplish a recursive search for each meaning (sense) till it arrives the text as i/p.
     - 3.1.2: Append the whole results to the suited Text-set.
4. Compute the weight of spam for suited employing weight of spam algorithm.
5. Get stopped.

The weight of SPAM algorithm calculates the weight of the SPAM for the suited-Text -set of the Algorithm of Recursive Text[3].

3.2.2 The Algorithm of Weight of the SPAM

Input : Suited Text-set

Process/Method : Compute the weight of SPAM employing the object in the suited Text-set

Output : The weight of SPAM

1. Get started
2. Start the weight of spam to "zero"
3. On the Text-set, for each and every text
   - 3.1: Choose the "HSN" referred as Highest Sense Number for the meanings
   - 3.2:Get updated the weight of spam by appending HSN as

     Weight of Spam=Weight of Spam+ HSN
4. Get stopped.

Each object in the Text-Set obtains from the pre-processing, has been took one by one and it compares against the Synset in the word-dictionary. Each and every function of the text as i/p has been appended to the matched text. The sense (meaning) number pretends as a counter variable and the Algorithm finds a specific word till it suits. The weight of the spam for the suited Text-set has been computed by employing the Algorithm of weight of spam.

The Figure 4 represents the "Synset" for the specific text as i/p "sport". In that "S" implies the "Synset", meant by, the semantic nexus (connection/relation) of specific text, while "W" shows the "word" meant by the lexical relation/connection. For the particular term "sport", the Word-net exhibits four meanings(senses) that have been denoted as S1,S2,S3,S4 andS5.

S1 : Athletics

S2 : mutant, mutation, variation, sport

S3 : summer cater

S4 : fun, play, sport

S5 : sportsman, sportswoman

Figure 4 Calculating the Sense number at Sample

The "Recursive Questing" performs in sensing of every 'Synset' till it gets reached the text as i/p. For an instance, "fun" has been one of the meanings of the term "sport" and the recursive-questing

built on the term "fun" till it reaches the "sport" in the Synset. This case has been shown at Figure 5. The greatest number of sensing detects from the suited Text Set's meanings and the suited text utilized for calculating the weight of spam. The meanings of the term "fun" have been implied as S1, S, S3 and S4.

S1 : fun, merriment, playfulness

S2 : fun, play, sport

S3 : Fun

S4 : playfulness, fun

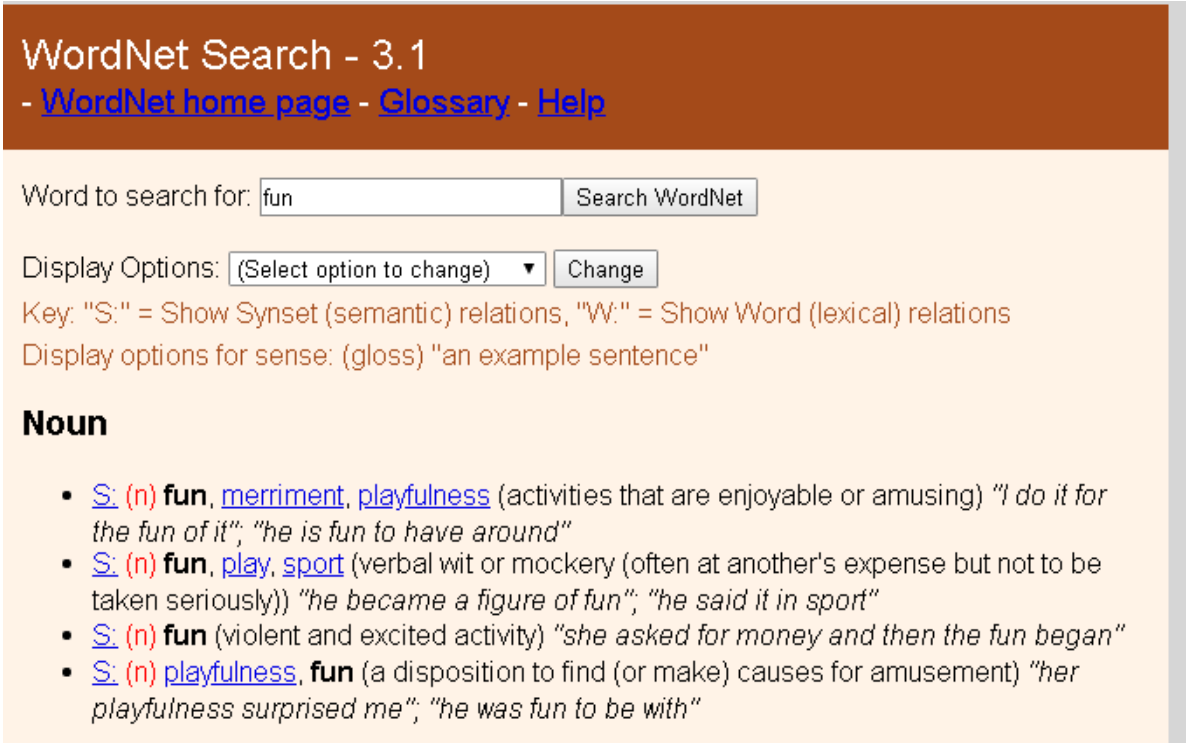


Figure 5 Recursive Search used for searching the yielded word.

3.3 Detecting the object/item and making mapped

At this level, this process/method has been extracted the contents of multimedia from the Multimedia-Set/a bunch of multimedia / a set of multimedia and gets detected. The item/objects represent over images employing the techniques of template-matching. The template-matching has been the technique of high-level machine-vision which recognizes the particles over an image which suits a template is a pre-defined one. The Algorithms based on the advanced –template matching; which access the template's finding occurrences / happenings nevertheless of its practices/demonstrations and the brightness belongs to local based[5].

The techniques of template matching have been so versatile and reliably, connectively straight forward to make utilized; that allows them to become one of eminent/immense familiar methods/processes of object localization. The power which is calculation one as the replica of biggie and completive templates can be a time-consumed due to the accessibility is utterly/literally so limited[6].

3.3.1 The Algorithm of Object Detection

Input : The multimedia set/the bunch of multimedia

Tool : Word net- dictionary

Process/Method: Recognize the object/item from the bunch of multimedia (multimedia-set) and calculate the weight of spam.

Output : Spam-weight suited object set

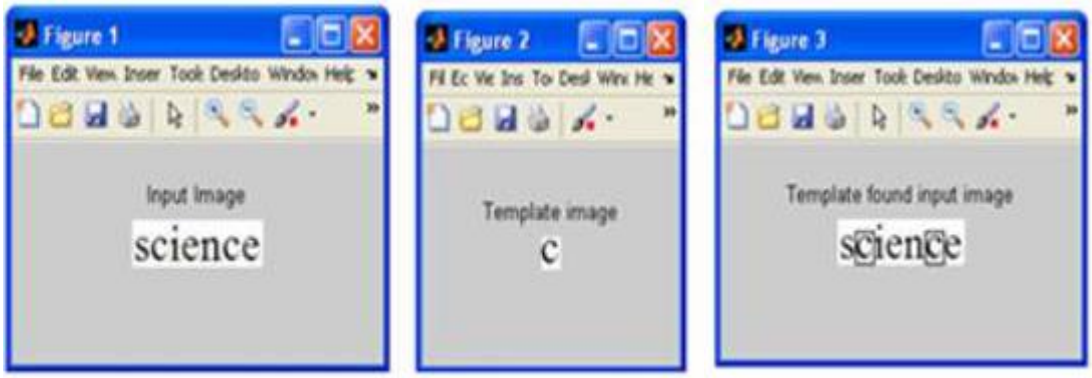
1. Get started
2. Reads the bunch of multimedia and recognizes the item/objects employing the technique of template –matching.
    - 2.1.1 Make compared the object/item with the word Dictionary employing the technique of template matching.
    - 2.1.2 Append the object/item into suited object set while the set of multimedia(multimedia-set) suits with the template-object-in the word net-dictionary
3. Compute the weight of spam for the suited object set employing the algorithm of spam-weight.
4. Get stopped.

By employing the technique of template matching, the process has a catalogue of words from mapping of an object. While the terms which are symbolic one, have been retrieved from the word-dictionary and calculate the object/item of spam weight that can be utilized for performing the spam – filtering at the stage/level of ending[7].

Here, the Set of Multimedia pretends as i/p. The o/p can be a Multimedia Object-Set (that secures the image set which received), uses for calculating the weight of spam. The technique of template-matching utilizes to suit the object of multimedia with an object/item on the WordNet dictionary[8].

Each and Every object/item from the suited object set has been read and compared the same with an object/item on the WordNet dictionary. The technique of template matching utilizes to make -

compared an object/item. Suppose the object/item a suit, the same kind of object/item appends to suited object-set.



(A) (B) (C)

Figure 6 some of the template matching instances.

The Figure 6 Vijayarani & Sakila (2015)[7] (A) refers an image as an i/p, (B) refers the template image on sample and (C)refers the suited template discovered from the image as an i/p one. It is an instance, of the matching of single letter. In the suggested algorithm, it has been focused on the particular text [9].

The Algorithm –the discussed one represented below, computes the object's spam-weight. The Spam-weight of object gets started with "0" (zero) to make sure that the earlier data was not stored at all. The Each and Every item/object represents at Suited Object Set, and then the object which is matched one, represents on the WordNet dictionary gets recognized. The numerous items recognized, have been appended to the sensing (meaning) number. The Object's spam weight computes employing HSN (Highest Sense Number) of each and every sense (meaning).

3.3.2 The Algorithm of Spam-weight for the suited object set

Input : Suited Object-set

Process/Method : Compute the weight of SPAM employing the object in the suited Object-set

Output : The Object weight of SPAM

1. Get started
2. Get started the Spam-weight to Zero
3. For each and every object/item in the set of objects

   3.1: Select HSN for sensing in the suited Object Set

   3.2: Get updated the weight of spam by appending HSN as

   Weight of Spam=Weight of Spam+ HSN
4. Get stopped.

3.4 The content based MFSC (Multi-Feature Spam Classifier)

At inception, the pre-processing performs gets followed by the calculation of textual feature approximation based on Recursive one and the calculation of spam-weight in the content based on MFSF (Multi Feature Spam Filtering) approachment. Detecting the object and making mapped have also been done, on the attachments to make computed the spam weight of object in the next level. The calculation of cumulative spam and the spam weight relies on the calculated weight of spam and the spam weight of object[8]. The classification of mails of spam has been occurred on the basis of cumulative spam that is calculated.

3.4.1 The Algorithm of Multi-Feature Spam Classifier

This Algorithm has been to help to make a decision on whether the mail receives as a genuine one (or) spam one. Attain the BLOB from the sets of Text Set/Multimedia Set. A BLOB has been a object that secures both objects such as multimedia & text[10]

The Spam weights cumulative calculate employing the BLOB. Suppose the spam weights cumulative increases the threshold of spam, while the received mail declares as a spam, if not, it deliberates as a genuine.

Algorithm for MFSC:

Input : The spam weight of object, the spam weight of text.

Process/Method : Compute the spam weight of object employing the object/item in the suited text set.

Output : Spam, genuine

1. Get started
2. Get initialized BLOB to Zero
3. Take BLOB as the spam weight of text and spam weight of object from pre-processing method/process
4. Calculate the weight of spam's cumulative employing the BLOB.
5. If the weight of spam's cumulative is greater than the threshold of spam, while,
   5.1: Classify the mails of spam
   else
   5.2:Classify the mails of Genuine
6. Get stopped.

3.4.2 The Spam-threshold setting

Based on the Scan results the Anti-spam tool decides a value that is a spam-threshold. The low level has been set, by default, on Anti-spam. Based on the set-threshold and the value, have been allocated a msg after the scan, then, the msg categorized as the spam or probable spam. The rules of spam-scanning have been employed to get set the threshold of spam on a mail server and a outlook[9]. The higher/big score, 'spam-like' msgs can be seen at exceed. If any msg scores 5 (or) beyond it .it gets held the pending-trap the rule of spam scanning embodies the rules of content-matching, DNS-based, check-sum based, statistical – filtering based. The value of threshold decides on the basis of rule of content – matching for this quest/research. The Table 1 represents the threshold chart on sample.

Table 1 Threshold chart on sample

Spam weight threshold	Maximum	High	Low	Minimum
Probable spam	60	70	80	90
Spam	75	85	90	100

Note:- The Threshold of spam weight has been allocated (or) rectified according to the tool requirements(merely, it decides whenever the mail scans for the SPAM)[10]

4. SUMMARY:

The Algorithm for intelligent multi-features approximation's tree structure spam classifier employing the big data for Emails is suggested. This process/method gets extracted the contents of both multimedia attachments and textual based from the emails at the stage of pre-processing. This process calculates the weight of spam by employing the technique of recursive textual-approximation while it accomplishes detecting the object and making mapped to calculate the spam weight of object. By employing two values, we can compute the spam weight's cumulative to make classified the E-mails. Moreover, this tool will be developed for calculating the weight of spam; therefore, the mail will get refused with the biggest weight of spam and also to enhance the ability of mail system. This research gets focused on the classification of E-mails and it deliberates the text content only on the E-mails and the attachments, template's matching as well. It deliberates as the ambits of this quest (research). In coming-days, this research might get elongated to assess the emails with attachments in terms of images; audio, videos etc. This research-basement offers the flourish-locations that helps to further research & researchers (pen-men) in the same domain.

REFERENCES

1. Yevseyeva, V, Basto-Fernandes, JR & Méndez, 'Survey on antispam single and multi-objective optimization', In Proceedings of International Conference on ENTERprise Information Systems, 2011;120–129.
2. Mu-Chun Su, Hsu-Hsun Loa & Fu-Hau Hsu, 'A neural tree and its application to spam e-mail detection', Expert Systems with Applications, 2010; 37: 7976–7985.
3. Shankar, S & Karypis, 'A Feature Weight Adjustment Algorithm for Document Categorization', In Proceeding of the 6th International Conference on Knowledge Discovery and Data mining (ACM SIGKDD), G 2000.
4. Ranganayakulu, Dhanalakshmi, L, Kavisankar, C & Chellapan, 'Enhanced E-mail authentication against spoofing attacks to migration phishing', European Journal of Scientific Research, 2011;54: 1.

5. Zhao, J, Basto-Fernandes, V, Jiao, L, Yevseyeva, I, Maulana, A, Li, R, Back, TA & Emmerich, ARXIV Computer Science "Multiobjective optimization of classifiers by means of 3-dconvex hull based evolutionary algorithm"[online],MTM 2014; https://www.arxiv.org/abs/1412.5710.
6. Yevseyeva, V, Basto-Fernandes, D, Ruano-Ordás, JR & Méndez, 'Optimising anti-spam filters with evolutionary algorithms', Expert System Applications, 2013; 40(10): 4010–4021.
7. Vijayarani,S & Sakila, 'Template Matching Technique for Searching Words in Document Images', International Journal on Cybernetics and Informatics, 2015; 4(6): 25–35.
8. Zhao, J, Basto Fernandes, V, Jiao, L, Yevseyeva, I, Maulana, A, Li, R, Back, T, Tang, K, Emmerich, MTM, 'Multiobjective optimization of classifiers by means of 3-D convex-hull-based evolutionary algorithms', Inf. Sci. 2016; 367–368: 80–104.
9. Zhou, B, Yao, YY & Luo, JG, 'Cost-sensitive three-way email spam filtering', Journal of Intelligent Information Systems, 2014; 42(1): 19–45.
10. https://en.wikipedia.org/wiki/WordNet.