# A Progressive Classification Framework for Detecting SPAM emails and Identification of Authors

**I V S Venugopal, D Lalitha Bhaskari, M N Seetaramanath**

***Abstract: Emails are the most popular form of communication in the space of cyber communications. In the recent past, many of the instances were observed, where the mode of communication were shifted to instance communication methods such as instance messages or video-based services for interaction. Nevertheless, for a detailed communication, there is no replacement of email communications. A number of surveys have reported that the amount of emails exchanged daily ranges between 200 to 250 million every day including the personal, business or promotional emails. Considering such a massive space for information exchange, it is regardless to mention that this space becomes the target for information misuses. One of the biggest threat to the email collaboration is spam emails containing unsolicited information or many of the cases asking for critical information of the recipients. Most of the email service providers help the users by incorporating a spam filtering process to prevent spamming in the email servers. Nonetheless, due to the critical nature of language used in communication makes the spam detection highly difficult. The fundamental strategies followed by most of the filters are to detect the spam emails based on specified key words. Regardless to mention, that in different domains of business or studies, some of the keywords carry different significance and cannot be blacklisted. Also, the inappropriate detection of the email as spam may lead to severe information loss. A good amount of research attempts is made in the recent past to build a framework for detection of spam as perfect as possible. However, due to the mentioned restriction the bottleneck still persists in between email filtration and detection of spam accuracy. Thus, this work proposes a novel automatic framework for detecting the spam emails on a wide range of domains. The obtained accuracy is significantly high for this framework due to the multiple layered approach adapted. The framework deploys classification of the emails in various domains and further applies the keyword-based filtration process with analysis of term frequency along with identification of nature of the sender for confirmation of the process resulting into progressive classification in order to make the world of email communication highly secure and satisfiable.***

***Keywords: Spam filtering, Term Frequency, Term Relation, Domain Knowledge, Author identification, progressive ,classification.***

## I. INTRODUCTION

The significance increases in the number of activities over internet, the increase of active users can be observed. In the due time the commonly known methods of communication were obsolete and users started finding a faster way of making the communications possible, thus the email communication came into existence. Today for a regular purpose user, it is observed that the number of email exchanges is ranging between 40 to 50 as per the report of R. Team [1]. The same report elaborates that, the number of emails for a business user can range between 100 to 150 per day and any business user has to spend a significant amount of time in processing the emails. It is to pragmatically identify that entire set of emails received or sent does not correspond directly to the business interest. Often the emails can contain information, which is unsolicited or promotional or an actual theft of information. Hence, in order to reduce the number of emails to work on a classification method for emails is a long-standing demand. The traditional methods of classifying emails are purely based on the text and as stated in this work, this existing method is not highly appropriate as the selection of texts in any email will differ from working domain of the email uses. Nonetheless, a number of research attempts have demonstrated the use of text classifiers for email classifications. The work by J. D. Brutlag et al. [2] has demonstrated the challenges faced by traditional classifiers for email classifications. Also, the work by W. W. Cohen et. al. [3] validates the same thought. Nevertheless, as a method email classification is widely accepted and the benefits cannot be ignored. Due to the wide acceptability of email classification, for a long time, classification of the emails is a dense area for researchers. The generic classifiers for email can segregate the emails into relevant to work, threat or phishing or SPAM. Any general purpose or generic email classification model must include a wide variety of classifiers and generate the classified email groups [Fig – 1].

**Revised Manuscript Received on March 20, 2019.**

**I V S Venugopal**, Department of IT, G V P College of Engineering (A), Visakhapatnam, Andhra Pradesh, India.

**Dr D Lalitha Bhaskari**, Department of CS&SE,AUCE(A),Andhra Pradesh, Visakhapatnam, India

**Dr M N Seetaramanath**, Department of IT, G V P College of Engineering (A), Visakhapatnam, Andhra Pradesh, India

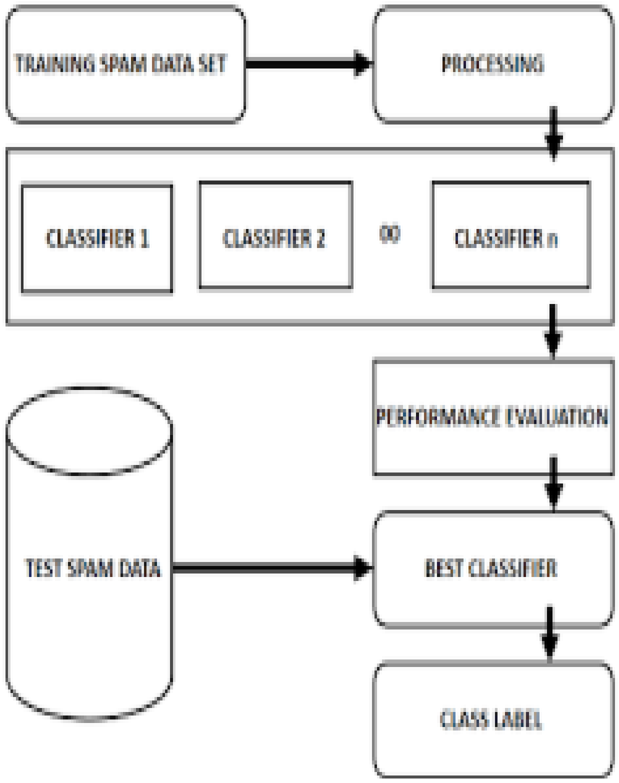


**Fig. 1 An Example of Generic Email Classifier**

A number of methods are deployed to achieve this classification purpose. One of the highly popular method for this purpose is the learning-based method such as the work by E. Blanzieri et al. [4]. The machine learning approaches have shown significant improvements over the traditional email classification methods in the recent past. The work by T. S. Guzella et al. [5] have compared the machine learning methods and showcased the advantages over other traditional methods. Continuing in the similar direction, S. Abu-Nimeh et al. [6] listed the machine learning methods for phishing detection. The most recent advancement in the space of spam or phishing email detection, the work of A. Almomani et al. [7] cannot be ignored, though it is highly argued for a similar method for detection with complete ignorance of the fact that domain specific content may fail in this method. Henceforth, it is natural to realize that the space of email classifications and detection of spam or phishing emails is highly diversified and the methods can be objected in the absence of domain specific keyword or knowledge bases. Thus, this work provides an automatic framework for detection of spam emails and authors based on domain specific term relations. The rest of the work is furnished as, in the Section – II the current updates in this field of research are listed, the email classification algorithm deployed in this framework is elaborated in the Section – III, the Section – IV elaborates on the proposed term discovery algorithm, the identification of author is formulated in the Section – V, in the Section – VI the complete workability of the framework is elaborated, further the obtained results are discussed in the Section – VII, in order to provide the knowledge of improvements the comparative analysis is presented in the Section – VIII and this work presents the final conclusion in the Section – IX.

## I. OUTCOMES OF THE PARALLEL RESEARCHES

The email classification has a wide range of applicability and a huge number of research attempts were made on this domain. In order to obtain better knowledge of this problem space, a detailed analysis is needed. Thus, in this section of the work, the outcomes from the parallel research attempts are reviewed and the shortcomings are identified. Identification of author or the nature of the email can be carried out successfully by identifying the characteristics or popularly known as features. The set of features plays a major role in identifying or separating each email or email author from other sets based on the values extracted for each email. The work by Y. W. Wang et al. [8] has showcased high accuracy of this strategy. Also, the work of M. R. Schmid et al. [9] in the similar line of progress, defines the benefits of customizable associative classification methods for feature and feature subset selection. The feature selection can also be applied for email texts in multiple languages. However, the pre-processing required for this method cannot be ignored as demonstrated by M. T. Banday et al. [10]. At times, the incorporation of features from different aspects of the email domains can expressively increase the efficiency. The notable work by M. Mohamad et al. [11] shows the advantages. Identifying the relations between the attributes or the features during the detection or classification process can also reduce the time complexity of the algorithms as suggested by N. A. Novino et al. [12] using graph-based methods. Apart from the feature selection methods, the supervised learning methods are also proven to be highly successful in detection of spam emails. The framework recommendations for building any such models are elaborated by W. Li et al. [13] emphasising the design aspects of the framework. These recommendations were well implemented by W. Meng et al. [14] and demonstrated the doles. In the machine learning approaches for detection of spam emails, the work by Z. J. Wang et al. [15] is also highly discussed for the benefits demonstrated and the notable strategy for weight assignments on various parameters. Finally, the summarization of the classification methods by S. A. Saab et al. [16] is highly appreciated and inferred in this work [Table – 1].

TABLE I
SUMMARY OF THE PARALLEL RESEARCH METHODS

| Method | Approach | Outcome | Identified Short Comings |
|---|---|---|---|
| M. R. Islam et al. [17] | Multi-Tier Classification | SPAM email detection | Domain specific key terms are ignored during the rule formation |
| A. A. Akinyelu et al. [18] | Random Forest | Phishing email detection | The availability of the multimedia data is ignored in the email texts |
| J. C. Gomez et al. [19] | PCA | SPAM and Phishing email detection | The extraction of features is limited to specific domain of communication and dependencies of the features are not identified |
| N. Al Fe'ar et al. [20] | Language Processor | Bi-Lingual email classification | The special symbols play a major role in multi lingual contents and the fact is overlooked |
| E. K. Jamison et al. [21] | Pairwise Classification | Thread classification | The association of the author and content is not highlighted |

Henceforth, the identified drawbacks are resolved in the proposed framework are explained in the subsequent sections of this work.

## II. AUTOMATED EMAIL CLASSIFICATION

The classification method used in this work is the term-based domain specific classification. As discussed in the previous sections of the work and validated by multiple research attempts, the domain specificity of the terms is highly significant for correct classification of the emails. Before elaborating the algorithm, this work lists the key words which can be considered as safe term for specific domain [Table – 2].

TABLE II
DOMAIN SPECIFIC SAFE TERM SUMMARY

| Domain | Term Analysis | | |
|---|---|---|---|
| | Identified Frequent Terms | Safe Terms | Term Relation |
| Finance | Additional income<br>Affordable new<br>Billing<br>Billion<br>Cash<br>Cheap<br>rates | Additional<br>Affordable<br>Billing<br>Cash<br>rates | <Extra, Added, Supplementary ><br><Reasonable, Inexpensive, Cheap><br><Promoting, Publicizing, Portraying><br><Money, Monies, Currency><br><taxes, charges, tariffs> |
| Education | Apply<br>Avoid<br>Be your<br>Certified<br>Congratulations<br>Compare<br>Score<br>Serious<br>Success | Apply<br>Avoid your<br>Certified<br>Congratulations<br>Score<br>Success | <Smear, Smear, Smear><br><Evade, Circumvent, Dodge><br><your, your, your><br><Expert, Specialized, Skilled><br><Cheers, Compliments, Felicitations><br><Marks, Value, Result><br><Achievement, Accomplishment, Feat> |
| Media and Advertisements | Buy<br>Call<br>free<br>Supplies<br>Refund<br>Remove<br>Request<br>Risk-free<br>Satisfaction | Call<br>free<br>Supplies<br>Refund<br>Satisfaction | <Noise, Song, Sound><br><allowed, permitted, welcome><br><Supplies, Supplies, Supplies><br><Reimbursement, Recompense, Compensation><br><Gratification, Consummation, Fulfillment> |
| News and Social Media | Cancel<br>Take<br>Terms<br>Trial<br>Unlimited<br>Urgent<br>Weight | Terms<br>Trial<br>Urgent | <Rapports, Relations, Standings><br><Experimental, Test, Pilot><br><Vital, Burning, Imperative> |

Further this work elaborates on the algorithm.

| **Algorithm - 1:** Automatic Email Classification |
|---|
| **Step - 1.** Accept the Initial Black Listed Terms |
| **Step - 2.** For each term |
| a. Build the term relation |
| b. Validate the terms for specified domain |
| c. Finalize the term black list |
| **Step - 3.** Accept the email corpus |
| **Step - 4.** Build the list of terms matching with term black list |
| **Step - 5.** For each term |
| a. Count the term frequency |
| b. If the term frequency > threshold |
| i. Mark the term as spam term |
| c. Count all spam terms |
| **Step - 6.** If the spam term frequency > threshold |
| a. Mark the email corpus as SPAM |
| **Step - 7.** Send the corpus for further validation |

Thus, as a result of the algorithm, the number of corpuses will be detected and will be sent to further validation by author characteristics. Regardless to mention that, the success of this algorithm highly depends on the term discovery for relevant fields for specified domains.

**The algorithm is visualized graphically here [Fig – 2].**

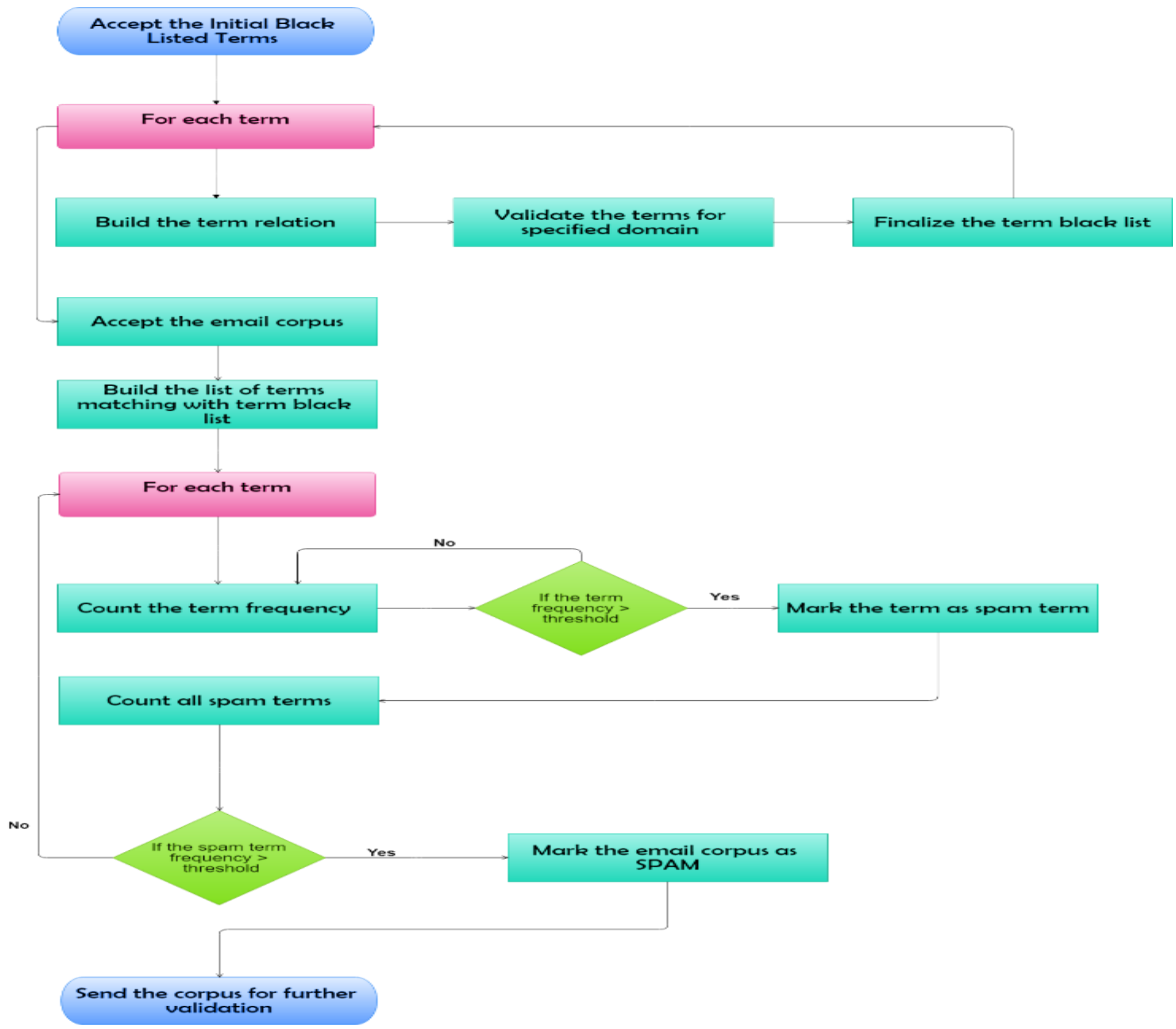


**Fig. 2 Proposed Email Classification**

The term discovery algorithm is discussed in the next section of this work.

## III. AUTOMATED TERM RELATION DISCOVERY

The term discovery plays a major role in this framework as the classifications of emails are dependent on the term-based classification. The relative term can be significantly beneficial for considering the safe terms and do not mark the email corpus as spam. For this purpose, finding the correct synonyms is the primary step. Hence, this work depends on the actual dictionary metadata for fetching the synonyms and further process the synonyms list with domain specific terms.

The proposed algorithm is furnished here.

**Algorithm - 2:** Term Relation Discovery

**Step - 1.** Accept the term list
**Step - 2.** for each term in the list
- a. Find the synonym for the term
- b. If the synonym belongs to domain term list
  - i. Calculate the relation score
  - ii. If relation score > threshold
    - 1. Then accept the term
- c. Else
  - i. Discard the term

**Step - 3.** Return relation list

**The algorithm is visualized graphically as well [Fig – 3].**

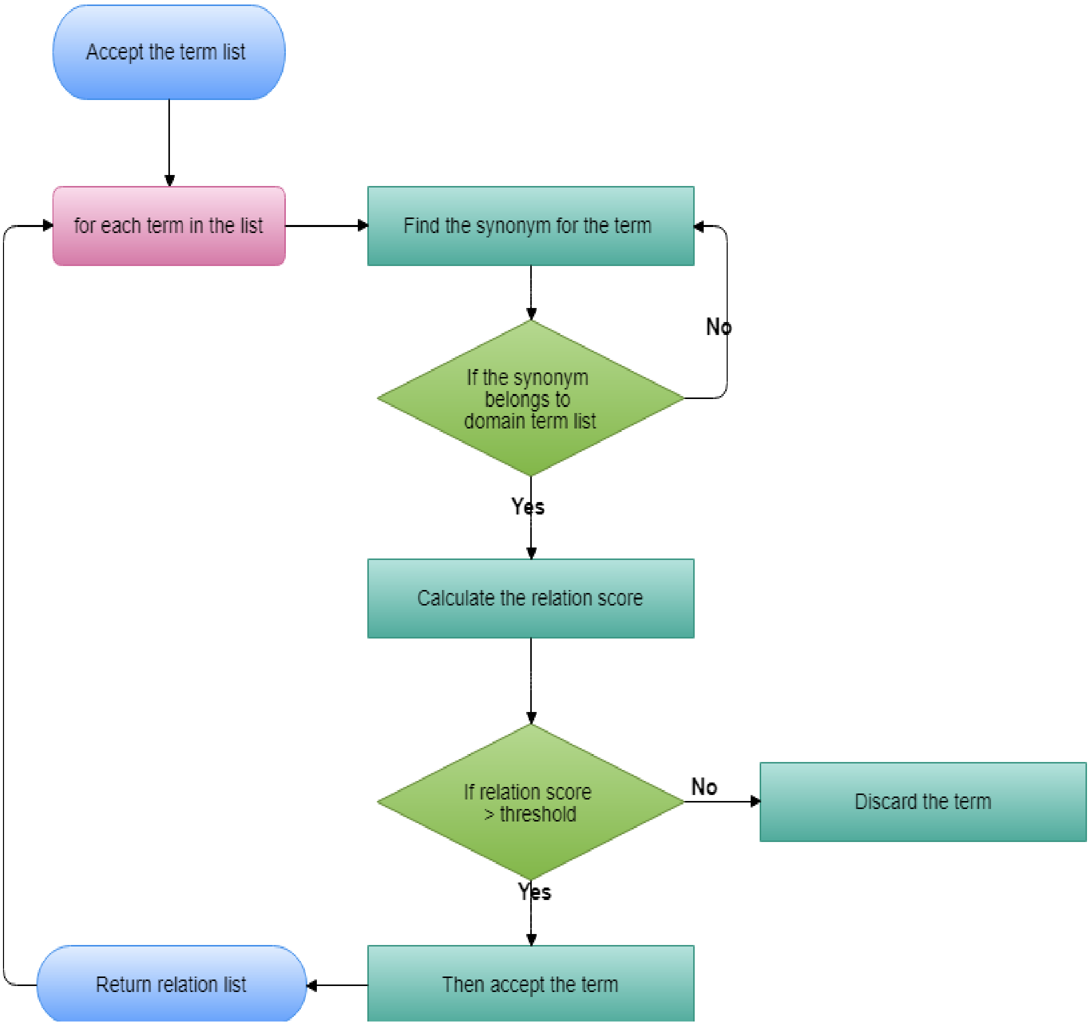


**Fig. 3 Proposed Automatic Term Discovery**

Henceforth, in the next section of the work, the author identification is elaborated.

## IV. AUTHOR IDENTIFICATION PROTOCOL

Further to the classification of email corpuses, the second level of validation is the author-based identification of the spam emails. In this section of the work, the identification protocol of the author is elaborated. Firstly, the description of the features of the author identification is listed here [Table – 3].

TABLE III
AUTHOR IDENTIFICATION PROTOCOL FEATURE LIST

| Feature Name | Feature Description | Possible Value Range |
|---|---|---|
| Author Email Domain | Domain of the email sender | Classified as public domain or private domain or corporate domain |
| Time Stamp | Time of the email received | Time Stamp |


Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

| | | |
|---|---|---|
| Email Size | Size of the email | KB |
| Attachments | The availability of the attachment in the email | 0 (No attachment)<br>Any Integer (Number of attachments) |
| Domain | Classification result of the email | Finance<br>Education<br>Media and Advertisements<br>News and Social Media |
| Safe Key words | Number of safe domain specific key words | Number |

Further the algorithm for author identification based on feature extract is elaborated here.

**Algorithm - 3**: Author Identification

**Step - 1.** Read the email with header
**Step - 2.** Separate the sender email address in "name" and "domain"
**Step - 3.** Switch case (domain)
: Public domain
: Private domain
: Corporate domain
**Step - 4.** Identify the time stamp of the email
**Step - 5.** Convert to local time stamp
**Step - 6.** Calculate the total email text size
**Step - 7.** Calculate the total email attachments size
**Step - 8.** Count the number of attachments
**Step - 9.** Apply key word search
**Step - 10.** Identify the domain of the email based on key words
**Step - 11.** Switch case (keyword list)
: Finance
: Education
: Media and Advertisements
: News and Social Media
**Step - 12.** Count the safe key words
**Step - 13.** Validate the author as SPAMMER or Not SPAMMER

The identification of the author helps in validation of email classification as the identification of the author and the email as spam can confirm the spam detection.

Further, in the next section of the work, the working flow of the entire framework is elaborated.

## V. Proposed Framework

The identification of email as spam can be controlled by analysing the email based on the key term-based classification, identification of domain specific terms, generation of term relation, identification of spam words, identification of the spam authors and finally validating the results with combination of knowledge from email and author classification or identification.[Fig - 4].

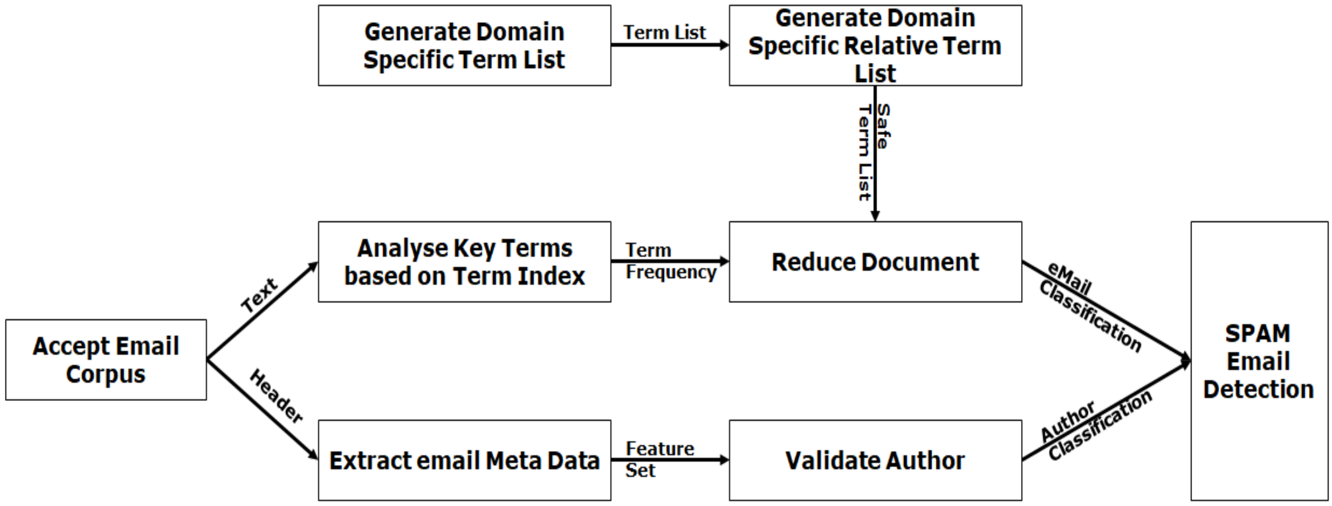


**Fig. 4 Proposed Framework**

The results obtained from this proposed framework are discussed in the next section of the work.

## VI. Results and Discussion

The results obtained from the proposed framework are highly satisfactory and cannot be deliberated without listing of the results. Thus, in this section of the work, the results obtained from each component are analysed and discussed. The results are furnished in five major components as initial classification results, discovery of the terms with domain specificity, classification or identification of the authors, final detection of spam emails and finally the performance of the complete framework.

### *A. Term Discovery Results*

Firstly, the term discovery results are analysed. The tern discovery phase, as elaborated in the algorithm, analyses the regular terms from the dictionary and performs synonyms extraction. Once the synonyms are extracted, then the domain specific terms and synonyms are extracted further. After the detection of list of domain specific term and synonyms, the lists of safe words are populated for each domain. The term discovery relations results are elaborated here [Table – 4].

TABLE IV
TERM RELATION RESULTS ARE EXTRACTED

| **Domain** | **Initial Terms** | **Number of Synonyms Generated** | **Domain Specific Terms** | **Domain Specific Safe Terms** |
|---|---|---|---|---|
| Finance | 97 | 5141 | 3599 | 1620 |
| Education | 253 | 12397 | 8678 | 3905 |
| Media and Advertisements | 333 | 13320 | 9324 | 4196 |
| News and Social Media | 180 | 10800 | 7560 | 3402 |

The results are visualized graphically here [Fig – 5].

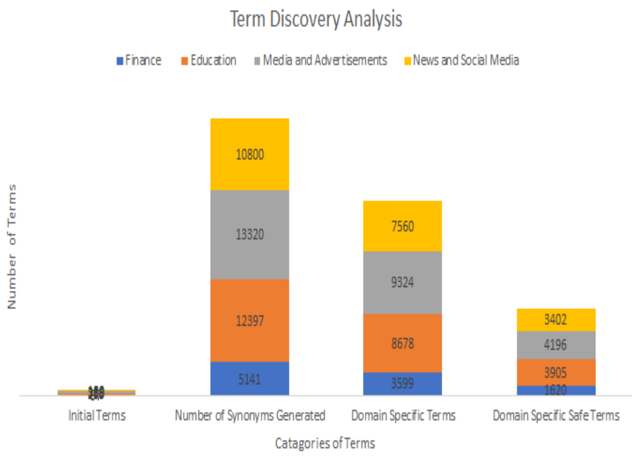


**Fig. 5 Term Discovery Analysis Results**

### *B. Initial Email Classification Results*

Secondly, the email classification results are discussed. The email corpus is provided to the framework and the frequency of spam terms are identified. Further the safe domain specific terms are reduced from the frequency list. Finally based on the decided frequency, that is 70% of the density of the words, the spam emails are identified.

The email classification results are elaborated here [Table 5].

**TABLE V**
**EMAIL CLASSIFICATION RESULT**

| **Corpus Name** | **Total Number of Words** | **Spam Words** | **Safe Words** | **Actual Spam Words** | **Threshold (70% Density)** | **Class** |
|---|---|---|---|---|---|---|
| corpus1.txt | 622 | 110 | 108 | 2 | 435 | Not SPAM |
| corpus2.txt | 176 | 160 | 5 | 155 | 123 | SPAM |
| corpus3.txt | 530 | 418 | 19 | 399 | 371 | SPAM |
| corpus4.txt | 310 | 101 | 100 | 1 | 217 | Not SPAM |
| corpus5.txt | 158 | 147 | 7 | 140 | 111 | SPAM |
| corpus6.txt | 724 | 531 | 28 | 503 | 507 | Not SPAM |
| corpus7.txt | 789 | 110 | 108 | 2 | 552 | Not SPAM |
| corpus8.txt | 101 | 97 | 3 | 94 | 71 | SPAM |
| corpus9.txt | 915 | 608 | 27 | 581 | 641 | Not SPAM |
| corpus10.txt | 576 | 435 | 20 | 415 | 403 | SPAM |
| corpus11.txt | 397 | 314 | 12 | 302 | 278 | SPAM |
| corpus12.txt | 716 | 110 | 108 | 2 | 501 | Not SPAM |
| corpus13.txt | 701 | 502 | 28 | 474 | 491 | Not SPAM |
| corpus14.txt | 171 | 157 | 4 | 153 | 120 | SPAM |
| corpus15.txt | 107 | 103 | 4 | 99 | 75 | SPAM |
| corpus16.txt | 422 | 107 | 105 | 2 | 295 | Not SPAM |
| corpus17.txt | 211 | 96 | 95 | 1 | 148 | Not SPAM |
| corpus18.txt | 906 | 602 | 30 | 572 | 634 | Not SPAM |
| corpus19.txt | 552 | 108 | 106 | 2 | 386 | Not SPAM |
| corpus20.txt | 606 | 110 | 108 | 2 | 424 | Not SPAM |
| corpus21.txt | 348 | 106 | 104 | 2 | 244 | Not SPAM |
| corpus22.txt | 850 | 110 | 108 | 2 | 595 | Not SPAM |
| corpus23.txt | 286 | 248 | 14 | 234 | 200 | SPAM |
| corpus24.txt | 968 | 621 | 24 | 597 | 678 | Not SPAM |
| corpus25.txt | 128 | 78 | 76 | 2 | 90 | Not SPAM |
| corpus26.txt | 531 | 110 | 108 | 2 | 372 | Not SPAM |
| corpus27.txt | 475 | 369 | 13 | 356 | 333 | SPAM |
| corpus28.txt | 174 | 88 | 86 | 2 | 122 | Not SPAM |
| corpus29.txt | 309 | 102 | 100 | 2 | 216 | Not SPAM |
| corpus30.txt | 375 | 320 | 11 | 309 | 263 | SPAM |

**The results are visualized graphically as well [Fig – 6].**

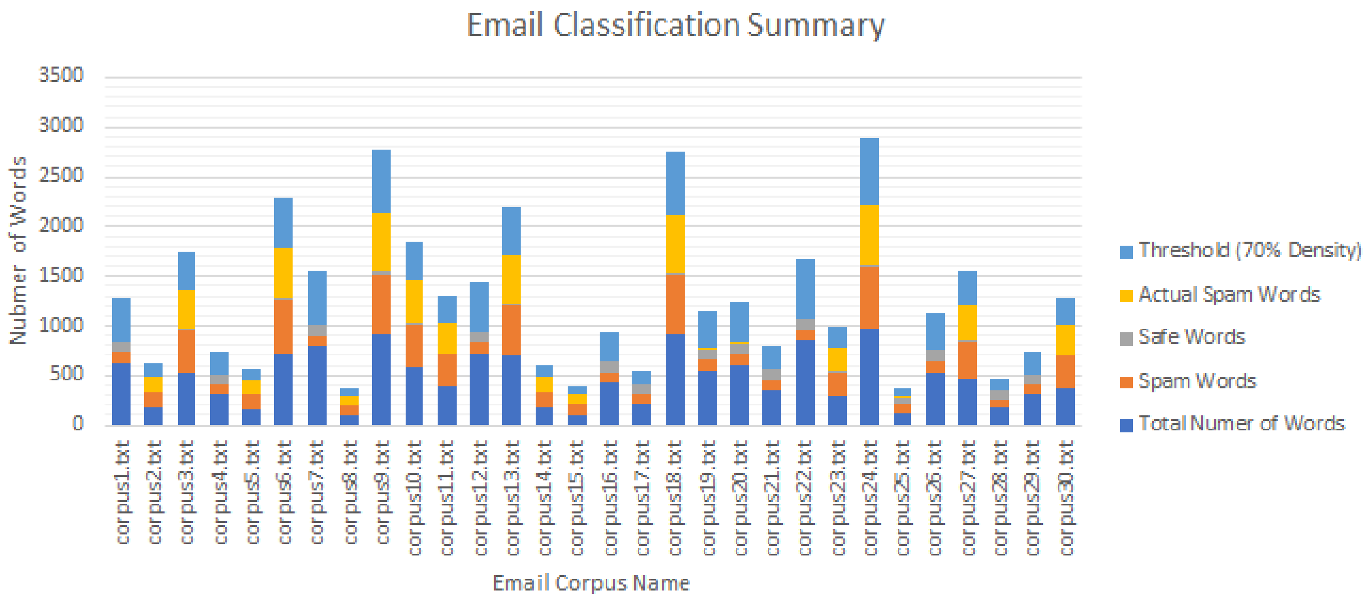


**Fig. 6 Email Classification Results**

### *C. Identification of Author*

Third, the identification of the author is valuable as based on the results of author identification, the final validation of the emails will be carried out.

The results from author identification phase are listed here [Table–6].

TABLE VI
AUTHOR CLASSIFICATION RESULT

| Corpus Name | Author Email Domain | Time Stamp | Email Size (KB) | Attachments | Domain | Safe Key words | Class (private domain and media or corporate domain and social) |
|---|---|---|---|---|---|---|---|
| corpus1.txt | public | 07:28:27 | 11196 | 0 | Edu | 108 | Not SPAMMER |
| corpus2.txt | corporate | 06:15:17 | 2464 | 0 | Media | 5 | Not SPAMMER |
| corpus3.txt | private | 06:29:24 | 9540 | 0 | Media | 19 | SPAMMER |
| corpus4.txt | public | 08:50:10 | 4340 | 0 | Media | 100 | Not SPAMMER |
| corpus5.txt | public | 08:56:44 | 3002 | 0 | Social | 7 | Not SPAMMER |
| corpus6.txt | corporate | 06:29:50 | 7240 | 0 | Social | 28 | SPAMMER |
| corpus7.txt | private | 06:16:34 | 12624 | 0 | Media | 108 | SPAMMER |
| corpus8.txt | private | 07:53:30 | 1818 | 0 | Media | 3 | SPAMMER |
| corpus9.txt | private | 07:48:50 | 14640 | 0 | Edu | 27 | Not SPAMMER |
| corpus10.txt | private | 07:15:37 | 6336 | 0 | Fin | 20 | Not SPAMMER |
| corpus11.txt | public | 07:27:06 | 7146 | 0 | Fin | 12 | Not SPAMMER |
| corpus12.txt | private | 06:23:17 | 13604 | 0 | Media | 108 | SPAMMER |
| corpus13.txt | corporate | 07:06:16 | 7711 | 0 | Fin | 28 | Not SPAMMER |
| corpus14.txt | public | 06:19:14 | 3249 | 0 | Media | 4 | Not SPAMMER |
| corpus15.txt | public | 07:38:08 | 1177 | 0 | Edu | 4 | Not SPAMMER |
| corpus16.txt | private | 08:00:18 | 4642 | 0 | Fin | 105 | Not SPAMMER |
| corpus17.txt | public | 07:58:22 | 3376 | 0 | Fin | 95 | Not SPAMMER |
| corpus18.txt | corporate | 06:51:51 | 12684 | 0 | Edu | 30 | Not SPAMMER |
| corpus19.txt | corporate | 08:55:59 | 5520 | 0 | Edu | 106 | Not SPAMMER |
| corpus20.txt | public | 07:27:37 | 9696 | 0 | Fin | 108 | Not SPAMMER |
| corpus21.txt | private | 08:38:50 | 5916 | 0 | Social | 104 | Not SPAMMER |
| corpus22.txt | corporate | 08:43:01 | 12750 | 0 | Edu | 108 | Not SPAMMER |
| corpus23.txt | private | 08:24:29 | 3146 | 0 | Media | 14 | SPAMMER |
| corpus24.txt | corporate | 06:05:37 | 9680 | 0 | Social | 24 | SPAMMER |
| corpus25.txt | corporate | 08:17:05 | 1536 | 0 | Social | 76 | SPAMMER |
| corpus26.txt | corporate | 06:33:30 | 9027 | 0 | Edu | 108 | Not SPAMMER |
| corpus27.txt | corporate | 06:44:23 | 5225 | 0 | Social | 13 | SPAMMER |
| corpus28.txt | public | 07:42:37 | 3306 | 0 | Edu | 86 | Not SPAMMER |
| corpus29.txt | corporate | 08:37:40 | 4635 | 0 | Media | 100 | Not SPAMMER |
| corpus30.txt | private | 07:42:40 | 4500 | 0 | Fin | 11 | Not SPAMMER |

### *D. Identification of SPAM Email as Progressive Classification*

Finally, the identification of spam emails is furnished here as the email must be identified as spam and the author of the same email also must be identified as spammer.

The final results are listed here [Table – 7].

TABLE VII
FINAL CLASSIFICATION RESULT

| Corpus Name | Email Class | Author Class | Spam Detection Result |
|---|---|---|---|
| corpus1.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus2.txt | SPAM | Not SPAMMER | Work Email |
| corpus3.txt | SPAM | SPAMMER | **Spam Email** |
| corpus4.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus5.txt | SPAM | Not SPAMMER | Work Email |
| corpus6.txt | Not SPAM | SPAMMER | Work Email |
| corpus7.txt | Not SPAM | SPAMMER | Work Email |
| corpus8.txt | SPAM | SPAMMER | **Spam Email** |
| corpus9.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus10.txt | SPAM | Not SPAMMER | Work Email |
| corpus11.txt | SPAM | Not SPAMMER | Work Email |
| corpus12.txt | Not SPAM | SPAMMER | Work Email |
| corpus13.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus14.txt | SPAM | Not SPAMMER | Work Email |
| corpus15.txt | SPAM | Not SPAMMER | Work Email |
| corpus16.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus17.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus18.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus19.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus20.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus21.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus22.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus23.txt | SPAM | SPAMMER | **Spam Email** |
| corpus24.txt | Not SPAM | SPAMMER | Work Email |
| corpus25.txt | Not SPAM | SPAMMER | Work Email |
| corpus26.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus27.txt | SPAM | SPAMMER | **Spam Email** |
| corpus28.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus29.txt | Not SPAM | Not SPAMMER | Work Email |
| corpus30.txt | SPAM | Not SPAMMER | Work Email |

Thus, it is natural to realize that, the identification of the spam emails is significantly narrowed down and considerably précised.

Further, the results from the corpus are elaborated here [Table – 8].

**TABLE VIII**
DATASET INFORMATION AND STATISTICS

| Dataset Description | Number of Emails (After Pre-Processing) | Number of SPAM Emails (After Pre-Processing) | Number of Authors | Number of SPAM Email Detected (By Proposed Framework) | Success (%) |
|---|---|---|---|---|---|
| **Title**: SPAM E-mail Database<br>**Donor**: George Forman<br>**Generated**: June-July 1999<br>**Modified**: April 2018 | 309 | 155 | 30 | 154 | 99.35 |

Hence, the success rate of detecting spam emails is highly satisfactory and it is to realize that, the success rate is achieved due to the incorporation of double classification of emails and authors.

### *E. Performance Analysis*

Additionally, the performance analysis of the framework is presented here [Table – 9].

TABLE IX

PERFORMANCE ANALYSIS

| Corpus Name | Time (MS) | Space (MB) |
|---|---|---|
| corpus1.txt | 1012 | 1.758331 |
| corpus2.txt | 10 | 4.177704 |
| corpus3.txt | 20 | 1.65506 |
| corpus4.txt | 11 | 4.482292 |
| corpus5.txt | 113 | 0.63073 |
| corpus6.txt | 116 | 2.187492 |
| corpus7.txt | 17 | 3.557358 |
| corpus8.txt | 18 | 3.93364 |
| corpus9.txt | 114 | 1.454681 |
| corpus10.txt | 810 | 3.008728 |
| corpus11.txt | 420 | 3.978645 |
| corpus12.txt | 114 | 1.375961 |
| corpus13.txt | 119 | 2.79998 |
| corpus14.txt | 18 | 3.407578 |
| corpus15.txt | 10 | 3.807327 |
| corpus16.txt | 12 | 4.634254 |
| corpus17.txt | 19 | 0.796867 |
| corpus18.txt | 20 | 2.55294 |
| corpus19.txt | 17 | 3.683876 |
| corpus20.txt | 16 | 0.917969 |
| corpus21.txt | 19 | 1.615593 |
| corpus22.txt | 15 | 2.909363 |
| corpus23.txt | 13 | 3.630646 |
| corpus24.txt | 116 | 1.289406 |
| corpus25.txt | 18 | 1.830376 |
| corpus26.txt | 15 | 2.711792 |
| corpus27.txt | 17 | 3.813362 |
| corpus28.txt | 12 | 4.301735 |
| corpus29.txt | 19 | 0.489326 |
| corpus30.txt | 10 | 2.603813 |

The result is visualized graphically as well [Fig – 7].

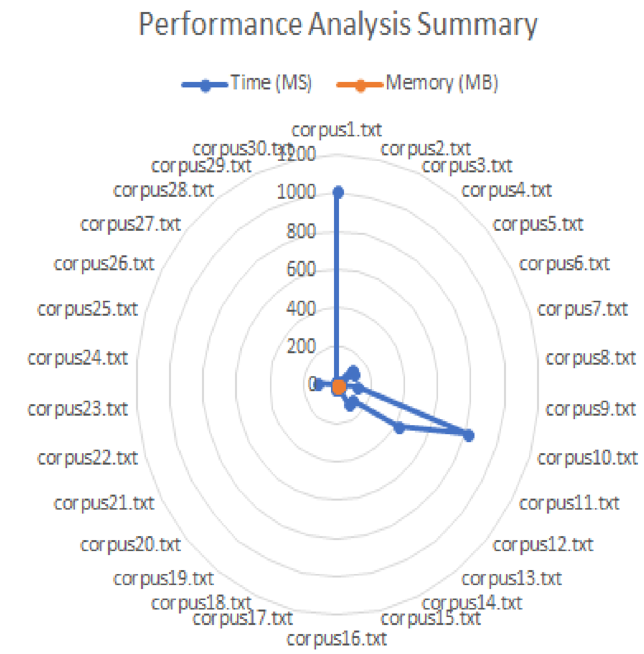


Fig. 7 Performance Analysis Result

## VII. COMPARATIVE ANALYSIS

In order to establish the thought of improvements over the existing methods, the comparative analysis must be carried out.

Thus, in this section of the work, the proposed framework is compared with the other parallel outcomes of the research [Table – 10] and ranked based on the factors such as functionalities like author detection, domain specificity and accuracy of detection.

TABLE X
COMPARATIVE ANALYSIS

| Method | Author Detection | Domain Knowledge | Accuracy | Rank (As High as Better) |
|---|---|---|---|---|
| J. Ratkiewicz et al. [22] | No | No | 90.91 | 4 |
| P.-A. Chirita et al. [23] | No | No | 89.95 | 3 |
| H. Yu et al. [24] | No | No | 85.94 | 1 |
| J. Ratkiewicz et al. [25] (Second Approach) | Yes | No | 84.9 | 2 |
| X. Hu et al. [26] | Yes | No | 95.89 | 5 |
| Proposed SIATR Framework | Yes | Yes | 99.35 | 6 |

Further, the accuracy analysis is also visualized graphically [Fig – 8].

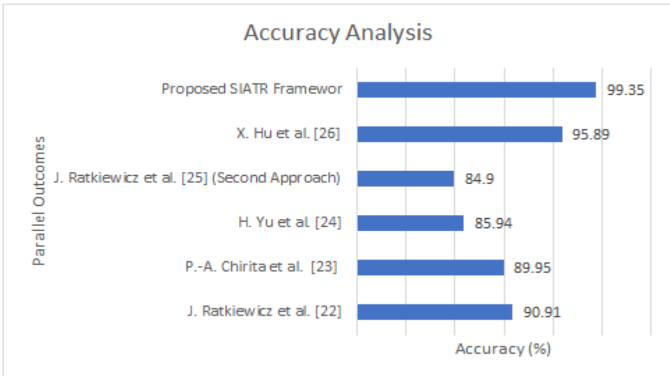


Fig. 8 Accuracy Analysis Result

Henceforth, with the understanding of the superiority of the proposed system compared with the other parallel methods, in the last section of this work, the final conclusion is presented.

## VIII. CONCLUSION

The importance of email communication in the field of education, research, corporate or personal communication cannot be ignored. The time taken for responding to each email is also significantly high for each individual and the fact of missing important communication cannot be ignored, thus this demands high time efficiency.

Also, this space of communication is also threatened by various malicious senders of emails as spam or never demanded information in form of advertisements or promotions or misleading information. Thus, the classification of emails as spam or work emails is deployed by various email service providers. Nevertheless, it is observed that many of the times, the actual work email is also classified as spam email, resulting into loss of information. Henceforth, this work proposes an automated framework for detection of spam based on domain specific knowledge, term-based information separation and finally based on the information about the authors. The proposed framework demonstrates high accuracy on real time and as well as on benchmark datasets. The multilevel verification and progressive classifications of the emails enable the least information loss and highly accurate detection of spam emails for making the world of email communication better, safer and more reliable.

## REFERENCES

1. R. Team, "Email statistics report 2015-2019", Mar. 2015.
2. J. D. Brutlag, C. Meek, "Challenges of the email domain for text classification", Proc. ICML, pp. 103-110, 2000.
3. W. W. Cohen, "Learning rules that classify e-mail", Proc. AAAI Spring Symp. Mach. Learn. Inf. Access, pp. 25, 1996.
4. Blanzieri, A. Bryl, "A survey of learning-based techniques of email spam filtering", Artif. Intell. Rev., vol. 29, pp. 63-92, Sep. 2008.
5. T. S. Guzella, W. M. Caminhas, "A review of machine learning approaches to spam filtering", Expert Syst. Appl., vol. 36, pp. 10206-10222, Oct. 2009.
6. S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, "A comparison of machine learning techniques for phishing detection", Proc. Anti-Phishing Work Groups 2nd Annu. Ecrime Res. Summit, pp. 60-69, 2007.

a. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, E. Almomani, "A survey of phishing email filtering techniques", IEEE Commun. Surveys Tuts., vol. 15, pp. 2070-2090, 4th Quart. 2013.

7. Y. W. Wang, Y. N. Liu, L. Z. Feng, X. D. Zhu, "Novel feature selection method based on harmony search for email classification", Knowl.-Based Syst., vol. 73, pp. 311-323, Jan. 2015.
8. M. R. Schmid, F. Iqbal, B. C. M. Fung, "E-mail authorship attribution using customized associative classification", Digit. Investigat., vol. 14, pp. S116-S126, Aug. 2015.
9. M. T. Banday, S. A. Sheikh, "Multilingual e-mail classification using Bayesian filtering and language translation", Proc. Int. Conf. Contemp. Comput. Informat., pp. 696-701, 2015.
10. M. Mohamad, A. Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification", Proc. 2nd Int. Conf. Comput. Commun. Control Technol., pp. 227-231, 2015.
11. N. A. Novino, K. A. Sohn, T. S. Chung, "A graph model based author attribution technique for single-class e-mail classification", Proc. 14th IEEE/ACIS Int. Conf. Comput. Inf. Sci. (ICIS), pp. 191-196, Sep. 2015.
12. W. Li, W. Meng, Z. Tan, Y. Xiang, "Towards designing an email classification system using multi-view based semi-supervised learning", Proc. 13th IEEE Int. Conf. Trust Secur. Privacy Comput. Commun. (TrustCom), pp. 174-181, Sep. 2015.
13. W. Li, W. Meng, "An empirical study on email classification using supervised machine learning in real environments", Proc. IEEE Int. Conf. Commun. (ICC), pp. 7438-7443, Jun. 2015.
14. Z. J. Wang, Y. Liu, Z. J. Wang, D. L. Liu, X. B. Zhu, K. L. Xu, D. M. Fang, "E-mail filtration and classification based on variable weights of the Bayesian algorithm" in Applied Science Materials Science and Information Technologies in Industry, Zürich, Switzerland:Trans Tech Publications Ltd, vol. 513, pp. 2111-2114, 2014.
15. S. A. Saab, N. Mitri, M. Awad, "Ham or spam? A comparative study for some content-based classification algorithms for email filtering", Proc. (MELECON), pp. 439-443, 2014.
16. M. R. Islam, J. Abawajy, M. Warren, Multi-Tier Phishing Email Classification with an Impact of Classifier Rescheduling, New York, NY, USA:IEEE, 2009.
17. A. Akinyelu, A. O. Adewumi, "Classification of phishing email using random forest machine learning technique", J. Appl. Math., vol. 2014, pp. 1-6, Apr. 2014.
18. J. C. Gomez, M. F. Moens, "PCA document reconstruction for email classification", Comput. Statist. Data Anal., vol. 56, pp. 741-751, Sep. 2012.
19. N. Al Fe'ar, E. Al Turki, A. Al Zaid, M. Al Duwais, M. Al Sheddi, N. Al Khamees, E-Classifier: A Bi-Lingual Email Classification System, New York, NY, USA:IEEE, 2008.
20. E. K. Jamison, I. Gurevych, "Headerless quoteless but not hopeless? Using pairwise email classification to disentangle email threads", Proc. 9th Int. Conf. Recent Adv. Natural Lang. Process., pp. 327-335, 2013.
21. J. Ratkiewicz et al., "Detecting and Tracking Political Abuse in Social Media", Proc. 5th Int'l AAAI Conf. Weblogs and Social Media, 2011
22. P.-A. Chirita, J. Diederich, W. Nejdl, "Mailrank: Using Ranking for Spam Detection", Proc. 14th ACM Int'l Conf. Information and Knowledge Management, pp. 373-380, 2005.
23. Yu et al., "Sybillimit: A Near-Optimal Social Network Defense against Sybil Attacks", IEEE/ACM Trans. Networking, vol. 18, no. 3, pp. 885-898, 2010.
24. J. Ratkiewicz et al., "Truthy: Mapping the Spread of Astroturf in Microblog Streams", Proc. 20th Int'l Conf. Comp. World Wide Web, pp. 249-252, 2011.
25. Hu et al., "Social Spammer Detection in Microblogging", Proc. 23rd Int'l Joint Conf. Artificial Intelligence, pp. 2633-2639, 2013.
26. Shivam Aggarwal,Vishal Kumar and S.D.Sudarshan,"Identification and Detection of Phishing Emails Using Natural Language Processing Techniques", Proceedings of the 7th International Conference on Security of Information and Networks,2014.
27. A. Pandian and Mohamed Abdul Karim, "Detection of Fraudulent Emails by Authorship Extraction",International Journal of Computer Applications (0975 – 8887), Volume 41– No.7, March 2012.
28. Hongming Che, Qinyun Liu and Lin Zou "A Content-Based Phishing Email Detection Method", IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C),2017.
29. H. Alghamdi, "Can Phishing Education Enable Users To Recognize Phishing Attacks" in Dublin Institute of Technology, Dublin, Ireland, 2017.