IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 17, 2017, accepted May 4, 2017, date of publication May 8, 2017, date of current version June 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2702187

Email Classification Research Trends: Review and Open Issues

GHULAM MUJTABA[1,2], LIYANA SHUIB[1], RAM GOPAL RAJ[3], NAHDIA MAJEED[2], AND MOHAMMED ALI AL-GARADI[1]

[1]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Department of Computer Science, Sukkur Institute of Business Administration, Sukkur 65200, Pakistan
[3]Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia

Corresponding authors: Ghulam Mujtaba (mujtaba@siswa.um.edu.my) and Ram Gopal Raj (ramdr@um.edu.my)

This work was supported by the University Malaya Research Grant under Grant RP026-14AET.

ABSTRACT Personal and business users prefer to use e-mail as one of the crucial sources of communication. The usage and importance of e-mails continuously grow despite the prevalence of alternative means, such as electronic messages, mobile applications, and social networks. As the volume of business-critical e-mails continues to grow, the need to automate the management of e-mails increases for several reasons, such as spam e-mail classification, phishing e-mail classification, and multi-folder categorization, among others. This paper comprehensively reviews articles on e-mail classification published in 2006–2016 by exploiting the methodological decision analysis in five aspects, namely, e-mail classification application areas, data sets used in each application area, feature space utilized in each application area, e-mail classification techniques, and the use of performance measures. A total of 98 articles (56 articles from Web of Science core collection databases and 42 articles from Scopus database) are selected. To achieve the objective of the study, a comprehensive review and analysis is conducted to explore the various areas where e-mail classification was applied. Moreover, various public data sets, features sets, classification techniques, and performance measures are examined and used in each identified application area. This review identifies five application areas of e-mail classification. The most widely used data sets, features sets, classification techniques, and performance measures are found in the identified application areas. The extensive use of these popular data sets, features sets, classification techniques, and performance measures is discussed and justified. The research directions, research challenges, and open issues in the field of e-mail classification are also presented for future researchers.

INDEX TERMS Email classification, spam detection, phishing detection, multi-folder categorization, machine learning techniques.

I. INTRODUCTION

With the increase in number of Internet users, email is becoming the most extensively used communication mechanism. In recent years, the increased use of emails has led to the emergence and further escalation of problems caused by spam and phishing emails. A typical user receives about 40–50 emails per day [1]; for others, hundreds of messages are usual. Users spend a significant part of working time on processing emails. Therefore, email management is an important issue faced by organizations and individuals, and it necessitates the need to devise mechanisms that intelligently classify and deal with the problem. Generally, the main tool for email management is automatic email classification [2], [3]. An automatic email classifier is a system that automatically classifies emails into one or more of a discrete set of predefined categories. For instance, for email management, one can benefit from a system that classifies an incoming email into official or personal, phishing or normal, and spam or ham.

Figure 1 shows the general architecture of automatic email classification. As shown in the figure, the email classification process is divided into three distinct levels: pre-processing, learning, and classification. To develop an automatic email classifier system, first, an email dataset should be collected. For example, if the aim is to develop an automatic spam email classifier, then one needs to collect a spam email dataset (i.e., the dataset containing both spam and non-spam used to train the classifier). Second, after data collection, the next task is to clean the dataset. The data cleansing task is generally known

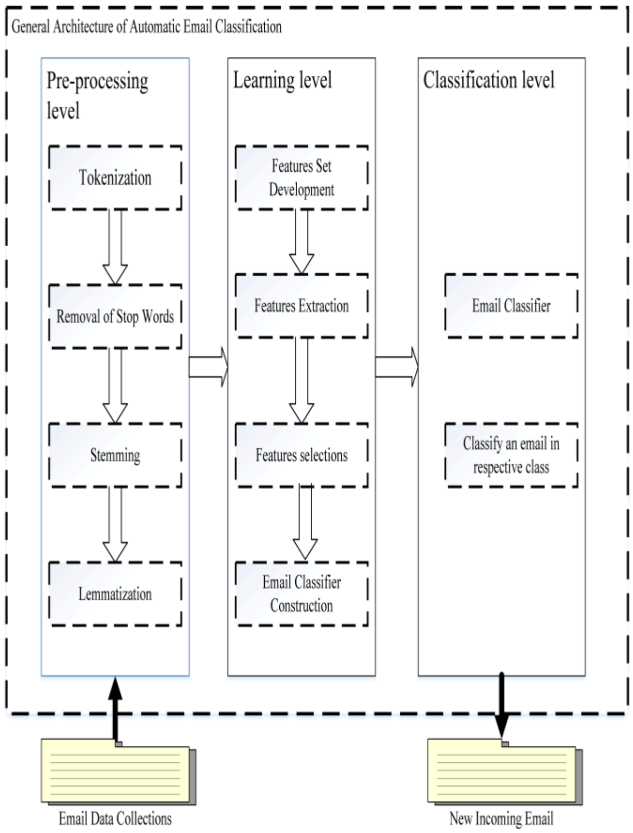


FIGURE 1. General architecture of automatic email classification.

as data pre-processing. In the pre-processing step, an email is converted into token of words. The pre-processing level also eliminates unnecessary words or stop words to reduce the amount of data that needs to be examined for their dispositions. Finally, in the pre-processing phase, stemming and lemmatization are performed on token of words to convert them into their root forms (e.g., "retrieving" to "retrieve"). At the learning level, features sets are developed and features are extracted. The term *feature* describes signs that represent a measurement of some aspect of a given user's email activity or behavior. In email classification, the effective extraction of a features set is essential in making the learning task efficient and more accurate. After feature extraction, the most discriminative features are selected for the classification to enhance classifier performance in terms of accuracy and efficiency. A classifier is constructed and saved to classify future incoming emails. Finally, at the classification level, a constructed classifier is used to classify an incoming email into a specific class, such as ham, spam, phishing, etc.

Currently, various experts are working in the email classification domain to classify an email into ham or spam or into phishing or legitimate. However, only a few review studies are available in the literature on spam email classification and phishing email classification from the text classification perspective. For example, Blanzieri and Bryl [4] presented a structured overview of existing learning-based approaches to spam filtering. In addition, a survey on datasets, text- and image-based features, performance measures, and spam filtering algorithms was presented. Guzella and Caminhas [5] investigated the available datasets, feature reduction techniques, and classification algorithms to identify spam emails. They also examined the literature on image-based spam email classification. Although both reviews are on spam email classification, they are outdated. After 2009, any review on spam email classification could not be found in the Web of Science and Scopus databases. To predict phishing emails, Abu-Nimeh et al. [6] compared the predictive accuracy of numerous machine learning algorithms, including logistic regression, classification and regression trees, support vector machines (SVMs), random forest, and neural networks. Almomani et al. [7] reviewed phishing email filtering techniques and presented the types of phishing attacks, phishing email classifications, and evaluation methods. However, the authors did not explore the publicly available datasets and various features for the detection of phishing email classifications.

Existing reviews reported either on spam or phishing email classification. Nonetheless, email classification is also used in other application areas, such as classifying an email into personal or official, complaint or non-complaint, and suspicious terrorist or normal, and classifying an incoming email into related directories, among others. Quality review articles should be produced to recapitulate the existing state-of-the-art email classification research for future researchers. Therefore, this study aims to conduct a comprehensive review on the application of email classification. The literature associated with the descriptors "Email Classification," "E-mail Classification," "Email Categorization," "E-mail Categorization," "Spam Email Detection," and "Phishing Email Detection" was collected comprehensively from the Web of Science and Scopus databases. Given the complexity and diversity of applications of email classification research, a methodological decision analysis framework for the selection of the collected articles was used. This framework targets the literature on five broad aspects: (1) email classification application areas, (2) email dataset analysis, (3) email features set analysis, (4) email classification technique analysis, and (5) performance measure analysis. This review comprised 98 studies from 2006 to 2016. This review can help researchers working in the field of spam email classification by answering following research questions:

(1) What are the various application areas where email classification has been applied?
(2) Which publicly available datasets can be accessed for the various application areas of email classification?
(3) What are the widely used features in the various application areas of email classification?
(4) What are the widely used machine learning techniques in the area of email classification?
(5) What performance evaluation metrics are employed to evaluate email classifier performance?
(6) What are the challenges and future research directions for future researchers working in the email classification domain?

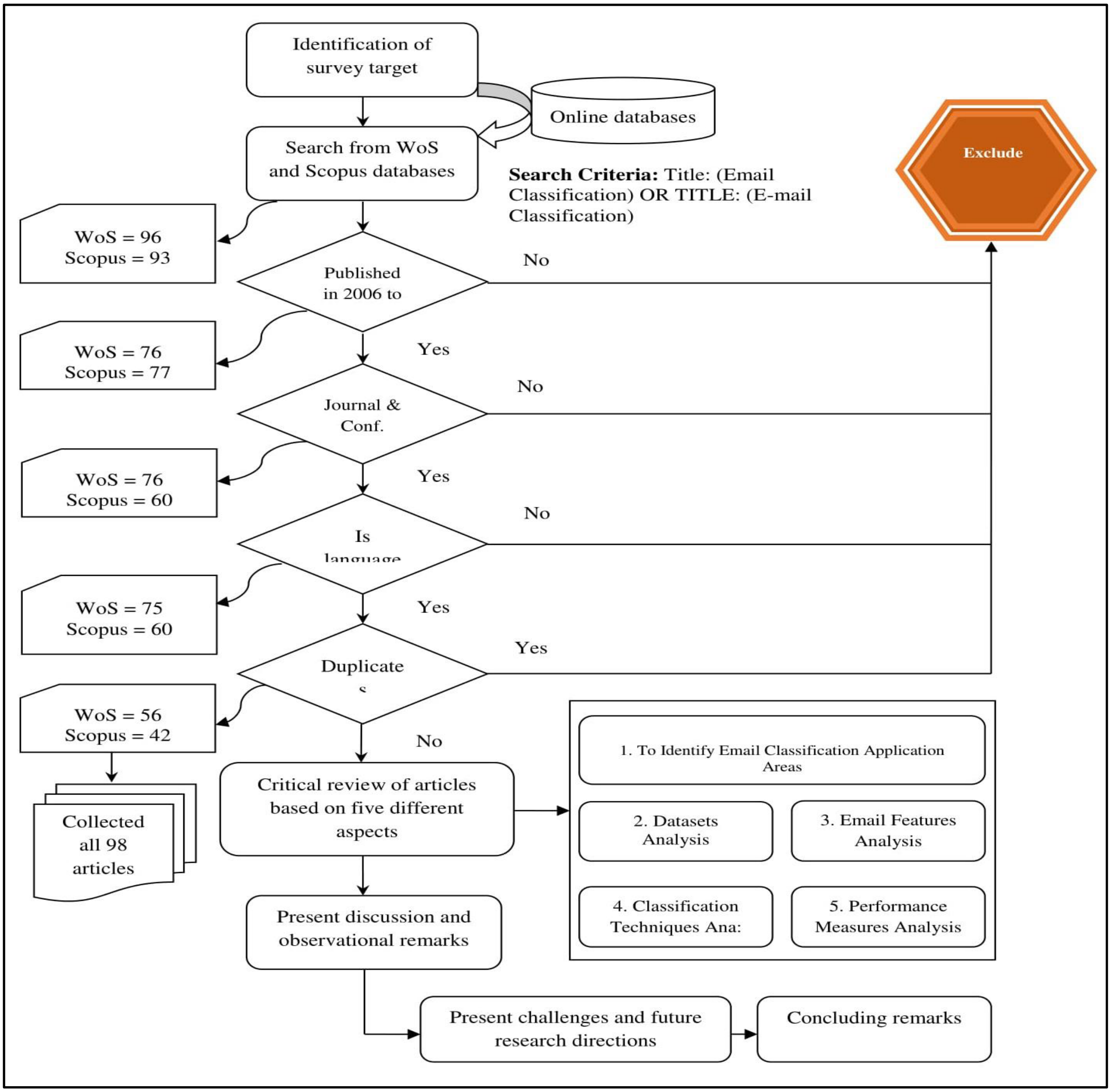

Identification of survey target
Online databases
Search from WoS and Scopus databases
Search Criteria: Title: (Email Classification) OR TITLE: (E-mail Classification)
Exclude
WoS = 96
Scopus = 93
Published in 2006 to
No
Yes
WoS = 76
Scopus = 77
Journal & Conf.
No
Yes
WoS = 76
Scopus = 60
Is language
No
Yes
WoS = 75
Scopus = 60
Duplicates
Yes
No
WoS = 56
Scopus = 42
Collected all 98 articles
Critical review of articles based on five different aspects
1. To Identify Email Classification Application Areas
2. Datasets Analysis
3. Email Features Analysis
4. Classification Techniques Ana:
5. Performance Measures Analysis
Present discussion and observational remarks
Present challenges and future research directions
Concluding remarks

FIGURE 2. Research Methodology.

The paper is organized as follows: Section 2 presents the research methods used for selecting the literature. Section 3 analyzes and discusses the categorical review of email classification research and gives the results. Section 4 presents some observations, open issues, and future research challenges. Section 5 concludes the paper.

II. RESEARCH METHODOLOGY

The research methodology of this review is illustrated in Figure 2. As previously mentioned, this study aimed to investigate holistically the research trends and patterns in the field of email classification. The following conditions were defined to limit the collection of articles:

(1) A comprehensive search was conducted. The articles were searched from the Web of Science and Scopus databases.

(2) The search strings for this review were "Email Classification," "E-mail Classification," "Email Categorization," "E-mail Categorization," "Spam Email Detection," and "Phishing Email Detection." The string-based search was performed on titles to retrieve the highly relevant articles on the topic under investigation.

(3) To report the latest trends in the application of machine learning techniques in email classification, only the studies that were published in 2006–2016 were used

for this review. The articles from 2006 were selected because this field became popular in that year.

(4) To achieve the highest level of relevance, international journal articles and conference proceedings were selected to represent comprehensively the related research communities. Thus, master's and doctoral dissertations, textbooks, unpublished articles, and notes were not considered for the investigation.

(5) Only articles published in the English language were extracted.

When the query was executed using the abovementioned search strings using the "title" field, 96 articles from Web of Science and 93 from Scopus were retrieved. Then, the year wise filter was applied to extract articles that were published in 2006–2016. The number of articles decreased to 76 from Web of Science and 77 from Scopus. The document type filter was then applied to retrieve the articles published either in international academic journals or in conference proceedings. This filter produced 76 articles from Web of Science and 60 from Scopus. Finally, the language filter was applied to select the articles that were published in the English language, and this filter produced 75 articles from Web of Science and 60 from Scopus. Duplicate articles that were present in both databases were removed. After removing the duplicate articles, 56 and 42 unique articles were extracted from Web of Science and Scopus, respectively. In sum, the five selection criteria produced 98 articles for this review. A comprehensive survey and analysis was performed on the selected 98 studies based on five aspects: (1) application areas, (2) datasets, (3) email features sets, (4) machine learning techniques, and (5) performance measures. The current trends, open issues, and research challenges were discussed in the email classification domain for future researchers.

III. EMAIL CLASSIFICATION STATE OF THE ART

This section presents a holistic analysis of email classification by assembling almost all major studies. The review can help researchers in this field to gain a better understanding of the existing solutions in the major areas of email classification. As discussed in the research methodology (Section 2), 98 articles were examined from five rationale aspects: email classification application areas, datasets used in application areas, features sets used in each application area, email classification algorithms, and performance measures. The review of all the rationale aspects is presented in Sections 3.1 and 3.5.

A. IDENTIFICATION OF EMAIL CLASSIFICATION APPLICATION AREAS

The review indicates that email classification is used in 15 application areas. These applications areas with the distribution of the number of studies are shown in Figure 3. For the sake of simplicity, these application areas are categorized into five domains: spam, phishing, spam and phishing, multi-folder categorization, and others, as shown in Figure 4. Other categories of the related application areas with only three or less studies, such as VIP email classification, business or personal email classification, and suspicious terrorist email classification, are included.

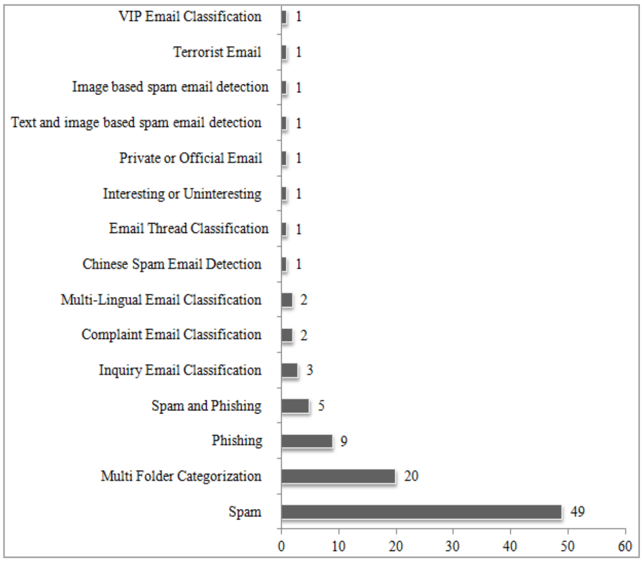


FIGURE 3. Distribution of articles according to applications area.

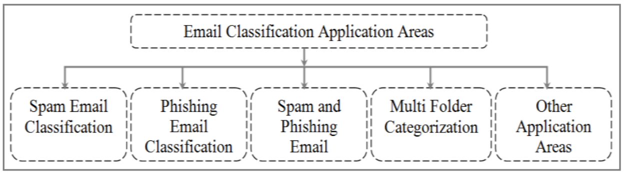

Email Classification Application Areas
Spam Email Classification
Phishing Email Classification
Spam and Phishing Email
Multi Folder Categorization
Other Application Areas

FIGURE 4. Application areas in email classification.

Figure 3 indicates that most studies on email classification are conducted to classify emails into spam or ham. Among the 98 articles, 49 are related to "spam email classification." Binary classifiers that classify emails into spam or ham were developed in the studies. The second highest number of articles is on the "multi-folder categorization of emails" (20 published articles), in which researchers developed a multi-class classifier that categorizes emails into various user-defined email directories. The third highest number of articles is related to "phishing email classification" (nine published articles), in which researchers developed binary classifiers that categorize emails into phishing or ham. The fourth highest number of articles is related to "spam and phishing email classification" (five published articles), in which researchers developed ternary classifiers that categorize emails into spam, phishing, or ham. Researchers recently classified spam email using text- and image-based features. A few researchers also developed techniques to classify emails into complaint or normal, inquiry or normal email, personal or email, interesting or uninteresting, VIP or normal email, and suspicious terrorist or normal email. The detailed distribution of the application areas with references is shown in Table 1.

TABLE 1. Distribution of articles according to application areas.

S. No.	Application Area	No. of Studies	References
1	Spam Email Classification	49	[8-56]
2	Multi Folder Categorization	20	[57-76]
3	Phishing Email Classification	9	[77-85]
4	Spam and Phishing Email Classification	5	[86-90]
5	Inquiry Email Classification	3	[91-93]
6	Complaint Email Classification	2	[94, 95]
7	Multi-Lingual Email Classification	2	[96, 97]
8	Chinese Spam Email Detection	1	[98]
9	Email Thread Classification	1	[99]
10	Interesting or Uninteresting Email Classification	1	[100]
11	Private or Official Email Classification	1	[101]
12	Text and image based spam email Classification	1	[102]
13	Image-based Spam Email Classification	1	[103]
14	Terrorist Email Classification	1	[104]
15	VIP Email Classification	1	[105]

B. EMAIL CLASSIFICATION DATASET ANALYSIS AND REVIEW

This section presents a detailed analysis of the datasets that were utilized in various application areas of email classification. Email classification is widely used in spam email classification, phishing email classification, spam and phishing email classification, and multi folder categorization of emails. Therefore, the researchers used public datasets to further explore and fine-tune these areas. The detailed analysis of the datasets used in various application areas is presented in Table 2.

Table 2 shows the application area of email classification, name of dataset, number of studies and their references (where a particular dataset is utilized), and total number of studies in a particular application area. The investigation reveals that the most popular dataset in spam email classification is the PU dataset. Out of the 49 studies on spam email classification, 10 used the PU dataset, followed by SpamBase dataset (eight studies), Enron spam email corpus (five studies), and SpamAssasin (five studies). The PU dataset is popular because the emails are derived from actual email messages sent to individuals. Moreover, the email messages are abstracted by replacing each distinct word with an arbitrarily chosen integer number, thus significantly reducing classification time and improved classification accuracy. A detailed comparative analysis of spam email classification datasets was also conducted [106].

The most widely used dataset in phishing email classification is PhishingCorpus [107]. This corpus includes 4,550 phishing email messages and does not include legitimate emails. Researchers utilized legitimate email subsets from existing spam email datasets for legitimate email messages. Out of the nine studies on the application area of phishing email classification, eight used phishing corpus along with the SpamAssasin dataset and one study adopted a custom dataset. PhishingCorpus is commonly used because it has a collection of hand-screened emails [108]. Moreover, the emails include different types of phishing targets and approaches, thus providing insights into the types of materials being sought. Researchers used PhishingCorpus in phishing and spam email classification for phishing emails and the combination of PU, LingSpam, SpamAssasin, TREC, and SpamBase datasets for spam detection.

Out of the 20 studies in the multi-folder email categorization, six used Enron dataset, one utilized TREC, and 13 adopted custom datasets. The Enron email dataset is widely used in the multi-folder categorization because it is the largest available dataset in email classification, with 252,757 preprocessed emails of 151 employees with 3,893 folders [109]. The Enron spam corpus should not be confused with the Enron email datasets. Enron spam email corpus is a successor of LingSpam, PU, and Enron email dataset. Details are provided in the literature [106]. Researchers in related areas of email classification used customized datasets. For instance, researchers on suspicious terrorist email classification developed and utilized "TCThreatening1" (which contains 500 terrorist emails, 481 spam emails, and 1,118 legitimate emails) and "TCThreatening2" (which contains 500 terrorist emails, 481 spam emails, and 2,912 legitimate emails) datasets [104]. Table 3 shows all the public datasets used in the application areas in email classification. The available links where the dataset can be downloaded and used are also shown.

C. FEATURE SET ANALYSIS AND REVIEW

Feature describes the properties that represent the measurement of some aspects of a given user's email activity or behavior. The extraction and selection of useful features in email classification are important steps to develop accurate and efficient classifiers. Researchers on email classification used the "bag of words" model, in which each position in the input feature vector corresponds to a given word or phrase. For example, the occurrence of the word "free" may be a useful feature in discriminating spam email. Therefore, carefully selected features can substantially improve classification accuracy and simultaneously reduce the amount of data required to obtain the desired performance. The features sets used in all the 98 studies on email classification are explored, as described in this section. The most widely used features

TABLE 2. Detailed analysis of datasets used in all identified areas of email classification.

Application Area	Name of Dataset	No. of Studies	Dataset Sample Size	Reference	Total
Spam Email Classification	PU	10	Total 7101 email (spam = 3020 and ham = 4081)	[13, 19, 33, 35, 40-43, 50, 55]	49
	Custom	9	It varies from study to study	[11, 18, 21, 23, 28, 45, 48, 52, 53]	
	SpamBase	8	Total 4601 emails (spam = 1813 and ham = 2788)	[14, 16, 20, 27, 29, 36, 44, 110]	
	Enron Spam Corpus	5	Total 30041 emails (spam = 13496 and ham = 16545)	[9, 12, 32, 34, 47]	
	SpamAssasin	5	Total 10744 emails (spam = 3793 and ham = 6951)	[24, 31, 46, 51, 54]	
	TREC	4	Total 92,189 emails (spam = 52,790 spam and ham = 39,399)	[10, 25, 30, 49]	
	CCERT	2	Total 34,360 emails (spam = 25,088 and ham = 9,272)	[15, 39]	
	LingSpam	1	Total 3252 emails (spam = 841 and ham = 2412)	[17]	
	Multiple: PU, Ling Spam, Enron, TREC	1	Discussed above	[8]	
	Multiple: LingSpam, Spam Assasin, TREC	1	Discussed above	[22]	
	Multiple: SpamAssasin, SpamBase	1	Discussed above	[26]	
	Multiple: SpamAssasin, TREC	1	Discussed above	[37]	
	Multiple: SpamAssasin, LingSpam	1	Discussed above	[38]	
Phishing Email Classification	Phishing Corpus with SpamAssasin	8	Total 11,501 emails (phishing emails = 4550 , ham emails from SpamAssassin = 6951)	[56, 77-84]	9
	Custom	1	Total emails 2034 emails (phishing = 1028 , ham = 1006)	[80]	
Spam and Phishing Email Classification	PU, LingSpam, SpamAssasin, TREC, Phishing Corpus	2	Phishing emails were taken from phishing corpus while spam and ham from respective spam classification datasets	[86, 87]	5
	Phishing Corpus, TREC	2	Phishing emails were taken from phishing corpus while spam and ham from respective spam classification dataset	[88, 89]	
	Phishing Corpus , SpamBase	1	Phishing emails were taken from phishing corpus while spam and ham from respective spam classification dataset	[90]	
Multi Folder Categorization	Enron Email Dataset	6	Sample size and categories varies from study to study	[57, 59, 63, 67, 69, 73]	20
	TREC	1	Total 92,189 emails in to 6 categories	[64]	
	Custom	13	Sample dataset and sample size varies from study to study	[61, 62, 65, 66, 68, 70-72, 74-76]	
Other Application Areas	Custom	15	Dataset and sample varies from study to study	[91-96, 98-105]	15

in email classification are email header, email body, email JavaScript, email URL, behavioral, SpamAssasin, network-based, Stylometric, term-based, offline, online, phrase-based, concept-based, rule-based, lexical, social, and structural features. The complete taxonomy of all these features based on the corresponding email classification application areas is shown in Figure 5. A brief overview of these features is presented as follows:

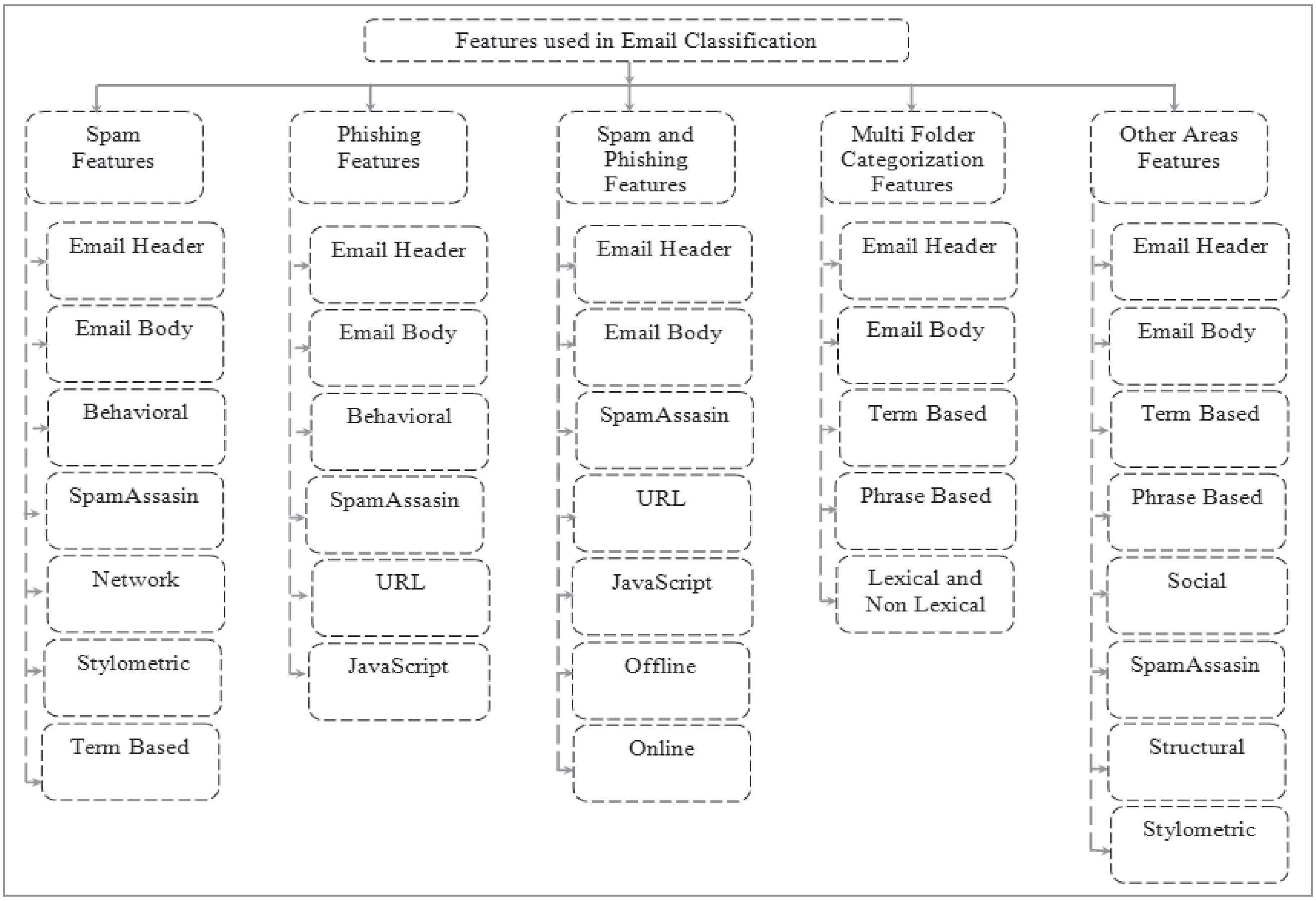


FIGURE 5. Feature set taxonomy used in email classification.

TABLE 3. List of publicly available datasets used in all five application areas with their available links.

S. No.	Dataset	Available Link
1	PU	http://www.csmining.org/index.php/pu1-and-pu123a-datasets.html
2	SpamAssasin	http://spamassassin.apache.org/publiccorpus
3	SpamBase	http://archive.ics.uci.edu/ml/datasets/Spambase
4	TREC	http://plg.uwaterloo.ca/~gvcormac/treccorpus07/
5	Enron	http://www.aueb.gr/users/ion/data/enron-spam/
6	CCERT	http://www.ccert.edu.cn/spam/sa/datasets. htm
7	LingSpam	http://www.csmining.org/index.php/ling-spam-datasets.html
8	PhishingCorpus	http://monkey.org/*jose/wiki/doku.php?id=PhishingCorpus

Email Header Features: Email header features are extracted and selected from an email's header. A header includes the from, to, bcc, and cc fields. For example, the popular email header features in phishing email classification are keywords, such as bank, debit, Fwd:, Re:, and verify in the subject field of an email. Other examples include the number of characters in the subject, number of words in the subject, word count in the from field, and non-model domain in the sender's email address.

Email Body Features: Email body features are selected from the email body part, which contains the main content of an email. Examples of email body features of the phishing email classification include HTML content in the body, HTML form in the body, dear keyword, number of characters and words, function words (e.g., credit, click, log, identify, information, etc.), suspension keyword, and verify your account keyword.

JavaScript Features: The JavaScript features include a JavaScript code in the email body. For example, the JavaScript features of the phishing email classification contain a JavaScript, OnClick event, pop-up window code, or any code in the email body that is loaded from an external website.

URL Features: URL features include suspicious URLs. Examples of URL features in the phishing email classification are the "@" sign in the URL, port numbers in the URL, presence of an IP address in the URL, number of URLs in the email body, when the URL has click, update, here, or login link text, or when the URL has two domain names.

SpamAssasin Features: SpamAssasin is an email filter that classifies emails into ham or spam. This intelligent email filter can identify spam using a diverse range of tests. Email headers and body are used in these tests to classify emails using advanced statistical methods. Its primary features are header tests, body phrase tests, Bayesian filtering, automatic address white list/black list, DNS block lists, and character sets, among others.

Offline Features: These features can be extracted locally and efficiently. Offline features are well suited for high-load context because these features must be handled in large mail servers. Examples include the number of pictures used as a link, non-ASCII characters in the URL, message size, and countries of links, among others.

Online Features: These features can be extracted online. Examples are OnClick event in the email, HTML form SSL protected, JavaScript status bar manipulation, and link domains being different from the JavaScript domain, among others.

Behavioral Features: These features can be used to determine atypical sending behavior. Examples include single email multinomial valued features such as presence of HTML, scripts, embedded images, hyperlinks, MIME types of file attachment, binary, or text documents, UNIX "magic number" file attachment, number of emails sent, number of unique email recipients, and number of unique sender addresses, among others.

Network-Based Features: Email features are extracted, selected, and aggregated on a per-packet basis to obtain the intra-packet score to be tagged to the email packet header. These features include packet size and TCP/IP headers, among others.

Stylometric Features: Stylometric features consist of the distinctive linguistic style and writing behavior of individuals to determine authorship. These features include the number of unique words, new lines, characters, function words, and attachments, among others.

Social Features: Social features consist of work-related and work-unrelated social relationships of employees during working hours. Examples of these features are domain name divergence, in-degree centrality of non-employee email recipients, occurrence ration of email recipients, occurrence ration of non-employee recipients, and cohesion of senders.

Structural Features: Structural features attempt to identify similar syntactic patterns between two texts while overlooking topic-specific vocabulary. Examples include pair of words occurring in the same order for two different emails.

Lexical and Non-Lexical Features: Non-lexical features are composed of descriptions of emails based on visual features (e.g., use of bold and capital letters or images), structural information (e.g., T field, CC, BCC, and abbreviations in the subject such as Fwd, Re, TR), characteristics of attachments (attached directly or included in a thread), and contextual information (presence of official signature and member of sender to the recipient social network). Lexical features include action authorization words (e.g., approve, request, please, thank you, to sign, etc.), action information (e.g., hello, possible, need, to provide, to transmit, to receive, etc.), action tasks (e.g., to discuss, to print, to share, must, follow up, etc.), action meeting (e.g., meeting, to post, periodically, etc.), and reaction tasks (e.g., to obtain, to relieve, to recruit, etc.).

Term-Based Features: The vocabulary list in term-based features is presented for classification. An incoming email is classified by term matching. Each term in a text pattern is described by a set of synonyms, generalizations, and specializations, among others.

Phrase-based Features: These features capture relevant phrases as a text pattern and not just a set of keywords. Phrase size can be fixed or variant.

Social Features: Social features include work related and work unrelated social relationships of employees during working hours. Examples of such features are: domain name divergence, in-degree centrality of non-employee email recipients, occurrence ration of email recipients, occurrence ration of non-employee recipients, and cohesion of sender.

Structural Features: Structural features attempt to identify similar syntactic pattern between two texts, while overlooking topic specific vocabulary. Examples include: pair of words occurring same order for two different emails.

Lexical and Non-Lexical Features: Non lexical features comprise of description of email based on visual features (such as use of bold, capital letters or images), structural information (such as T field, CC, BCC, abbreviations in subject such as Fwd, Re, TR), characteristics of attachment (attached directly or included in a thread), and contextual information (presence of official signature, member of sender to recipient social network). While lexical features include: action authorization words (such as approve, request, please, thank you, to sign, etc.), action information (such as hello, possible, need, to provide, to transmit, to receive, etc.), action tasks (such as to discuss, to print, to share, must, follow-up, etc.), action meeting (such as meeting, to post, periodically, etc.) and reaction tasks (such as to obtain, to relieve, to recruit, etc.).

Term Based Features: In term based features, list of vocabulary is prepared for classification. An incoming email is classified by term matching. Each term in text pattern described by set of synonyms, generalization, specialization, etc.

Phrase Based Features: These features capture relevant phrases as a text pattern not just a set of keywords. Phrase size may be fixed or variant.

Table 4 to Table 8 show the email features used all identified application areas. The most widely used features in all application areas of email classification are email header features and email body features. Nevertheless, behavioral and SpamAssasin features are also essential and useful in spam email classification. A possible reason is that "from field," "to field," "bcc field," and "subject field" of email

TABLE 4. Features used in spam email classification.

S. No.	Features used in application area of Spam email classification							Reference
	Header	Body	Behavioral	SpamAssasin	Network	Stylometric	Term Based	
1	✓	✓	X	X	X	X	X	[53]
2	✓	✓	X	X	✓	X	X	[54]
3	✓	✓	X	X	X	X	X	[48]
4	✓	✓	X	X	X	X	✓	[49]
5	✓	✓	X	X	X	X	X	[50]
6	✓	✓	X	✓	X	X	X	[51]
7	✓	✓	X	X	X	X	X	[52]
8	✓	✓	X	X	X	X	X	[41]
9	✓	✓	X	X	X	X	X	[42]
10	✓	✓	X	✓	X	X	X	[42]
11	✓	✓	X	X	X	X	X	[44]
12	✓	✓	X	X	X	X	✓	[45]
13	✓	✓	X	X	X	X	X	[46]
14	✓	✓	X	✓	X	X	X	[47]
15	✓	✓	X	✓	X	X	X	[35]
16	✓	✓	X	X	X	X	X	[36]
17	✓	✓	X	X	✓	X	X	[37]
18	✓	✓	X	X	X	X	X	[38]
19	✓	✓	X	X	X	X	X	[55]
20	✓	✓	X	X	X	X	X	[39]
21	✓	✓	X	X	X	X	X	[40]
22	✓	✓	✓	✓	X	X	X	[33]
23	✓	✓	✓	X	X	X	X	[34]
24	✓	✓	X	✓	X	X	X	[30]
25	✓	✓	✓	X	X	X	✓	[31]
26	✓	✓	✓	X	X	X	✓	[32]
27	✓	✓	X	✓	X	X	X	[24]
28	✓	✓	✓	X	X	X	X	[25]
29	✓	✓	✓	X	X	X	X	[26]
30	✓	✓	X	X	X	X	X	[27]
31	✓	✓	X	X	X	X	X	[28]
32	✓	✓	X	X	X	X	X	[29]
33	✓	✓	X	X	X	X	X	[21]
34	✓	✓	X	X	X	X	X	[22]
35	✓	✓	X	X	X	X	X	[23]
36	✓	✓	X	X	X	X	X	[15]
37	✓	✓	X	X	X	X	X	[16]
38	✓	✓	X	X	X	X	X	[110]
39	✓	✓	X	X	X	X	X	[17]
40	✓	✓	X	✓	✓	X	X	[18]
41	✓	✓	✓	X	X	X	X	[19]
42	✓	✓	✓	✓	X	X	X	[20]
43	✓	✓	X	X	X	X	X	[8]
44	✓	✓	X	X	X	✓	X	[9]
45	✓	✓	X	X	X	X	X	[10]
46	✓	✓	X	X	X	X	X	[11]
47	✓	✓	X	X	X	✓	X	[12]
48	✓	✓	X	X	✓	X	X	[13]
49	✓	✓	✓	✓	X	X	X	[14]

headers in spam emails may contain the most powerful features for the identification of spam email. Moreover, an email body may include some discriminative features in classifying email into spam or ham. SpamAssasin features are specially designed to detect spam emails. Therefore, these features enhance the accuracy of spam email detection classifiers. Behavioral features, such as the number of unique email recipients and the number of unique sender addresses, are also

TABLE 5. Features used in phishing email classification.

S. No.	Features used in application area of phishing email detection						Reference
	Header	Body	URL	JavaScript	Behavioral	SpamAssasin	
1	✓	✓	X	X	X	X	[80]
2	✓	✓	X	X	X	X	[56]
3	✓	✓	X	X	X	X	[79]
4	✓	✓	X	✓	✓	X	[81]
5	✓	✓	✓	✓	X	✓	[82]
6	X	✓	✓	✓	X	X	[77]
7	✓	✓	✓	✓	X	X	[83]
8	✓	✓	X	X	X	X	[84]
9	✓	✓	✓	✓	X	X	[78]

TABLE 6. Features used in spam and phishing email classification.

S. No.	Features used in application area of phishing and Spam email detection							Reference
	Header	Body	URL	JavaScript	SpamAssasin	Offline	Online	
1	✓	✓	X	X	X	X	X	[90]
2	✓	✓	✓	✓	✓	✓	✓	[88]
3	✓	✓	✓	✓	✓	✓	✓	[89]
4	X	✓	X	X	X	X	X	[86]
5	X	✓	X	X	X	X	X	[87]

discriminative features used to classify spam emails in many studies [31]–[34].

URL and JavaScript features are the most frequently used features in the phishing email classification, and they significantly improve the accuracy of phishing email classifiers [77], [78], [82], [83]. This result may be due to most phishing emails containing either suspicious URLs that may redirect to unidentified and suspicious Web pages or form fields that may require some sensitive information to fill and submit. Header, body, and term-based features are imperative in multi-folder categorization and other application areas [92]–[94], [99], [100] because emails can be automatically classified in a predefined category based on the terms used in the email body part and email ‘‘subject’’, ‘‘to’’, or ‘‘from’’ fields.

D. REVIEW AND ANALYSIS OF TEXT CLASSIFICATION TECHNIQUES

Email classification techniques are classified into five different categories: supervised machine learning, unsupervised machine learning, semi supervised machine learning, content-based learning, and statistical learning [45], [111]. The classification is illustrated in Figure 6. The learning algorithm in supervised machine learning is provided with input instances, and output labels do not easily identify a function that approximates this behavior in a generalized manner. Examples of supervised learning techniques are SVM, decision trees, genetic algorithm, artificial neural network, Naive Bayes, Bayesian network, and random forest.

The learning algorithm in unsupervised machine learning is provided with input instances but with output labels and learning algorithm attempts to identify similar patterns in the input instances to determine an output. One example is clustering using K-means. Semi-supervised machine learning is actually a supervised machine learning technique using small labeled data and is not a supervised machine learning method that requires large labeled data. This approach is conducted by using some labeled data to facilitate a classifier in labeling unlabeled data. One example is active learning. Content-based email classification techniques use keywords in emails for classification. Statistical learning techniques assign a probability or score to each keyword, and the overall score or probability is used to classify incoming emails.

Researchers on email classification used all types of techniques, but among them, supervised machine learning is the most commonly used. Figure 7 shows the distribution of email classification techniques in all the application areas. Supervised machine learning is the most widely used technique among all the listed methods. Out of the 98 studies, 71 used supervised learning, 14 used content-based techniques, 9 adopted statistical techniques (direct statistical properties of the class), 2 used unsupervised machine learning techniques, and 2 utilized semi-supervised machine learning techniques. An overview of the email classification

TABLE 7. Features used in multi folder categorization.

S. No.	Features used in application area of multi folder categorization							Reference
	Header	Body	Term Based	Phrase Based	Concept Based	Rule Based	Lexical	
1	✓	✓	✓	X	X	X	X	[76]
2	✓	✓	✓	X	X	X	X	[75]
3	✓	✓	X	X	✓	X	X	[74]
4	✓	✓	X	X	X	X	X	[97]
5	✓	✓	X	X	X	X	X	[70]
6	✓	✓	X	✓	X	X	X	[71]
7	✓	✓	X	X	X	✓	X	[72]
8	✓	✓	X	X	X	X	X	[73]
9	✓	✓	X	✓	X	X	X	[65]
10	✓	✓	✓	X	X	X	X	[68]
11	✓	✓	X	X	X	X	X	[67]
12	✓	✓	X	X	X	X	X	[66]
13	✓	✓	X	X	X	X	X	[69]
14	✓	✓	✓	✓	X	X	X	[94]
15	✓	✓	✓	X	X	X	X	[64]
16	✓	✓	✓	X	X	X	X	[61]
17	✓	✓	X	X	X	X	✓	[62]
18	✓	✓	X	X	X	X	X	[63]
19	✓	✓	X	X	X	X	X	[58]
20	✓	✓	✓	X	X	X	X	[57]

techniques is presented in Table 9. The table is grouped according to type of email classification. Each row contains the technique name and the number of studies in spam classification, phishing, spam and phishing, multi-folder categorization, and other application areas. SVM is the most frequently used technique in supervised machine learning (17 out of 71 studies), followed by decision trees (9 out of 71 studies), Naive Bayes (7 out of 71 studies), K-nearest neighbor (5 out of 9 studies), and random forest (4 out of 71 studies). Only 2 out of the 98 studies used semi-supervised machine learning, and both studies adopted different techniques, that is, voting algorithm with active learning and SVM with active learning. Two studies adopted unsupervised techniques. The authors in both studies used the K-means clustering technique. Nine out of the 98 studies used statistical learning. The most frequently used technique in statistical learning is ranking method (2 out of 9 studies). Furthermore, 14 out of the 98 studies used content-based learning. The most frequently used technique in content-based learning is simple-term statistics (5 out of 14 studies), followed by concept-based learning and language code similarity (2 out of 14 studies).

The above analysis shows that supervised machine learning techniques are widely used to classify emails. Moreover, the accuracy of these techniques is significant because supervised machine learning techniques are easy to use and learn from the training data. Prediction accuracy increases when training data are extensive. The only issue with supervised machine learning is the labeling of training data. Data labeling may entail a long period of time. SVM is widely used in supervised machine learning techniques, and it provides more accurate results than other techniques. SVM can produce better results with all the available features in the master feature vector because it is not prone to over-fitting [112]. After SVM, decision trees are the second most frequently used classifier to categorize emails. This approach also showed promising results in numerous studies. The performance of J48 in email classification is good because it does not require domain knowledge and it can deal with high-dimensional data. Moreover, J48 can handle datasets with errors and missing values. It is considered a nonparametric classifier, which means that it does not use assumptions in the classifier structure. The main disadvantage of J48 is that it can easily overfit.

E. REVIEW AND ANALYSIS OF PERFORMANCE MEASURES

Classifier accuracy can be determined by calculating the number of true positive class cases (TP), true negative class

TABLE 8. Features used in other related application areas of email classification.

S. No.	Features used in other related application areas of email classification								Reference
	Header	Body	Term Based	Phrase Based	Social	SpamAssasin	Structural	Stylometric	
1	✓	✓	✓	X	X	X	X	X	[93]
2	✓	✓	X	X	X	X	X	X	[105]
3	✓	✓	✓	X	X	X	X	X	[100]
4	✓	✓	X	X	X	X	X	X	[97]
5	✓	✓	X	X	X	X	X	X	[103]
6	✓	✓	X	X	X	X	✓	✓	[104]
7	✓	✓	X	X	X	X	X	X	[98]
8	✓	✓	✓	X	X	X	X	X	[92]
9	✓	✓	X	✓	X	X	X	X	[91]
10	✓	✓	✓	✓	X	X	X	X	[94]
11	✓	X	X	X	✓	X	X	X	[101]
12	✓	✓	X	X	✓	X	X	X	[60]
13	✓	✓	✓	X	X	X	✓	✓	[99]
14	✓	✓	X	X	X	X	X	X	[58]
15	✓	✓	X	X	X	✓	X	X	[102]

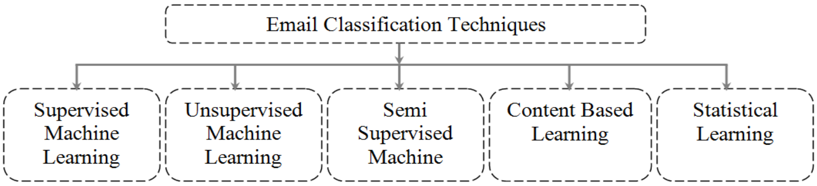

Email Classification Techniques
Supervised Machine Learning
Unsupervised Machine Learning
Semi Supervised Machine
Content Based Learning
Statistical Learning

FIGURE 6. Types of email classification techniques.

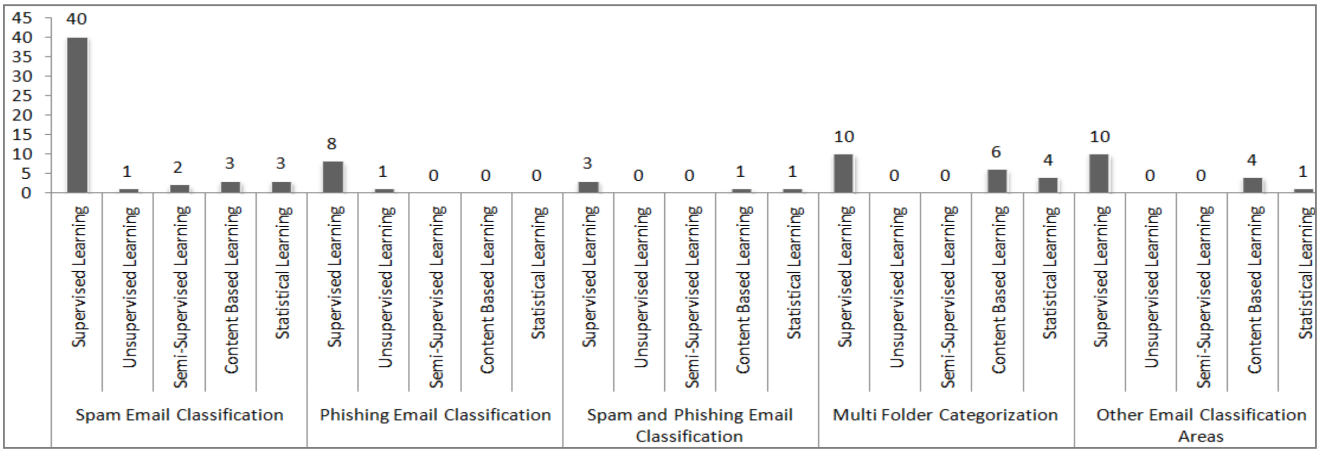


FIGURE 7. Frequency of email classification techniques.

cases (TN), false positive class cases (FP), and false negative class cases (FN). These numbers form a confusion matrix, as shown in Table 10 for binary classification problems. TP is the rate of correctly classified instances that belong to the "yes" class and is predicted as "yes." TN is the rate of correctly classified instances that belong to the "no"' class and is predicted as "no." FP is the rate of incorrectly classified instances that belong to the "yes" class and is

TABLE 9. Summary of email classification techniques.

S. No.	Technique Name	Spam Classification	Phishing Classification	Spam and Phishing Classification	Multi Folder Categorization	Others	Total
	Supervised Machine Learning Technique						
1	Support Vector Machines	11	2	1	2	1	17
2	Decision Tree	5	0	1	1	2	9
3	Naïve Bayes	3	1	0	1	2	7
4	K-Nearest Neighbor	1	0	0	3	1	5
5	Random Forest	1	3	0	0	0	4
6	Artificial Neural Network	2	0	0	1	0	3
7	Rough Set Theory	2	0	1	0	0	3
8	AdaBoost	1	0	0	0	1	2
9	Artificial Immune Based System	1	0	0	0	1	2
10	Bayesian Network	1	1	0	0	0	2
11	Ad Infinitum	0	0	0	0	1	1
12	Case Base Reasoning	1	0	0	0	0	1
13	Deep Belief Network	0	0	0	1	0	1
14	Fuzzy Support Vector Machine	1	0	0	0	0	1
15	Hybrid: (K-Means and SVM)	1	0	0	0	0	1
16	Hybrid: Decision Tree and Naïve Bayes	1	0	0	0	0	1
17	Hybrid: DT and Ontology	1	0	0	0	0	1
18	Hybrid: DT+NB	0	0	0	1	0	1
19	Hybrid: Naive Bayes, AdaBoost and Support Vector Machines	1	0	0	0	0	1
20	Hybrid: Naive Bayes, AdaBoost and Support Vector Machines	1	0	0	0	0	1
21	Hybrid: Rough Set and Support Vector Machine	1	0	0	0	0	1
22	Hybrid: Support Vector Machine, AdaBoost and Naïve Bayes	1	0	0	0	0	1
23	Improved Naïve Bayes	1	0	0		0	1
24	Logistic Regression	0	0	0	0	1	1
25	PART	0	1	0	0	0	1
26	Particle Swarm Optimization	1	0	0	0	0	1
27	proposed Consultation Algorithm	1	0	0	0	0	1
	Semi-Supervised Machine Learning Technique						
29	Voting algorithm using Naive Bayes, IBK, J48 and SMO with Active Learning	1	0	0	0	0	1
30	Support Vector Machines with Active Learning	1	0	0	0	0	1
	Unsupervised Machine Learning Technique						
31	K-Means Clustering	1	1	0	0	0	2
	Statistical Learning						
32	Dynamic category hierarchy	0	0	0	1	0	1
33	Granular Classification and Load Sensitive Classification	0	0	0	1	0	1
34	Incremental Forgetting Weighted Bayesian	1	0	0	0	0	1
35	Nonnegative Matrix Factorization (NMF)	0	0	1	0	0	1
36	Partial Memory Incremental Learning Algorithm FLORA2	1	0	0	0	0	1
37	Quadratic-Programming Linear Model	0	0	0	0	1	1
38	Ranking Method	1	0	0	1	0	2
39	Rule Based	0	0	0	1	0	1
	Content Based						
40	Ant Colony Based Spam Filter	1	0	0	0	0	1
41	Concept Based Statistics	0	0	0	2	0	2
42	Dictionary based approach	0	0	0	0	1	1
43	Language Code Similarity	0	0	0	1	1	2
44	OCR Filter, VBOW filter, Cascade Filter	0	0	0	0	1	1
45	Ontology based learning	1	0	0	0	0	1
46	PCADR	0	0	1	0	0	1
47	Simple Term Statistics	1	0	0	3	1	5

predicted as “no.” FN is the rate of incorrectly classified instances that belong to the “no” class and is predicted as “yes.”

Sokolova and Lapalme [113] presented the most widely used performance measures for binary classification and multi-class classification, as shown in Table 11.

TABLE 10. Confusion Matrix.

		Actual Class	
		Yes	**No**
Predicted class	**Yes**	TP	FN
	No	FP	TN

The typical performance indices of two distinct areas are commonly used, namely, information retrieval (recall, precision, and F measure) and decision theory (false positives and false negatives), to evaluate the performance of an email classifier. Table 12 presents the frequency distribution of the performance measures based on the application areas. The most frequently used performance measures in the application areas of spam and phishing are accuracy, recall, precision, f-measure, false positive rate, and false negative rate. Precision, recall, f-measure, and accuracy are the most widely used performance measures in the multi-folder categorization and other application areas.

Researchers utilized accuracy, precision, recall, and f-measure as performance metrics to evaluate the performance of a classifier. However, these metrics alone are insufficient to evaluate classifier performance correctly. Various studies have indicated that the dataset, as well as the number of emails in the datasets, is imbalanced. For example, PhishingCorpus was utilized to classify emails into normal or phishing [56], [77]–[84]. The number of phishing emails was lower than that of normal emails. Moreover, LingSpam dataset was utilized to classify emails into spam and ham [17]. Here, the number of spam emails in the dataset is also lower than that of ham emails. Therefore, in cases in which the dataset is imbalanced, the most suitable performance metric is area under the curve [114], [115]. This metric is appropriate in evaluating the performance of a classifier with respect to a specific class. Moreover, researchers proposed a ternary email classifier to categorize emails into "ham," "spam," or "phishing" and used simple precision, recall, or accuracy to evaluate the performance of a classifier [86]–[90]. The suitable measures for multi-class classifiers using an imbalanced dataset are macro precision, macro-recall, and overall accuracy [113].

IV. FUTURE RESEARCH DIRECTIONS

This section highlights several research challenges and open issues in the current studies on email classification. In this regard, substantial research work is yet needed to enhance the performance of email classification in its various application areas. These research are presented below.

(1) This section highlights several research challenges and open issues in the current studies on email classification. In this regard, substantial research is needed to improve the performance of email classification in its various applications.

(1) Use of ontology and semantic web: Experts can focus on classifying emails using ontology. Moreover, classified results can be used in semantic web by creating a modularized ontology based on the classified result. An adaptive ontology can be planned and created as an email filter based on the classification result. This ontology can be developed and customized on the basis of a user's report when the user requests a suspicious email report. By creating a suspicious email filter in the form of an ontology, a filter can be customized, scalable, and modularized by a user and thus be embedded in other systems.

(2) Real-time learning (stream learning) of email classifier: Majority of existing research on email classification is based on datasets that may not include real-time environmental elements. Only one [14] out of the 98 studies performed real-time testing of email classifiers. Real-time environmental factors greatly affect the performance of email classifiers. Therefore, the performance of email classifiers in real environments can be evaluated, as an online stream of emails is more complicated that an offline dataset. The evaluation of email classifiers in real time remains a potential research challenge for experts. In addition, email users who use email classification services are crucial, and the usability of the classifiers can be tested based on their experience.

(3) Dynamic updating of the feature space: Another area of research is designing methods that enable the incremental addition or removal of features without re-building the entire model to keep up with new trends in spam or phishing email classification.

(4) Deep learning: Deep learning enables computational methods with several processing layers to learn representations of data with several levels of abstraction [116], [117]. The main characteristic of deep learning is that the layers of features are not human engineered. These features are learned from data using a general-purpose learning process that changes the feature-engineering task from human-engineered features to automatic engineering features. These algorithms are useful in email classification with high-dimensional data, in which human-engineered features do not effectively reflect the learning vectors from given data.

(5) Email classification using hierarchical classification: For email classification with varying granularity, such as email classification with sub-categorization, classifiers must distinguish among several email characteristics to calculate the final classification. To facilitate these processes, complex classification issues may be solved by breaking them down into several smaller classification tasks in which classifiers are prepared in a hierarchy. The first classifier in this classification calculates a high-level classification, for example, whether an email is spam or not, and low-level classifiers are trained for different sub-classes of the high-level classification. The low-level classifiers used for the specific classification of initial classifiers thoroughly distinguish the different types of spam emails and their danger levels.

TABLE 11. Performance measures for binary class and multi-class classification [113].

S. No	Measure	Classification Type	Formula	Evaluation Focus
1	Precision	Binary	$\Pr ecision = \frac{tp}{tp+fp}$	Overall effectiveness of a classifier
2	Recall	Binary	$\operatorname{Re} call = \frac{tp}{tp+fn}$	Class agreement of the data labels with the positive labels given by the classifier
3	F-Measure	Binary	$Specificity = 1/2(\frac{tp}{tp+fn} + \frac{tn}{tn+fp})$ $F - Measure = \frac{(\beta 2+1)tp}{(\beta 2+1)tp+\beta 2 fn+fp}$	Effectiveness of a classifier to identify positive labels
4	Accuracy	Binary	$Accuracy = \frac{(tp+tn)}{(tp+fn+fp+tn)}$	Relations between data's positive labels and those given by a classifier
5	Specificity	Binary	$Specificity = \frac{tn}{fp+tn}$	How effectively a classifier identifies negative labels
6	AUC	Binary	$Specificity = 1/2(\frac{tp}{tp+fn} + \frac{tn}{tn+fp})$	Classifier's ability to avoid false classification
7	Average Accuracy	Multi-Class	$\sum_{i=1}^{l}(\frac{tpi+tni}{tpi+fni+fpi+tni})/l$	The average per-class effectiveness of a classifier
8	Error Rate	Multi-Class	$\sum_{i=1}^{l}(\frac{fpi+fni}{tpi+fni+fpi+tni})/l$	The average per-class classification error
9	Precision$_\mu$	Multi-Class	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fp_i)}$	Agreement of the data class labels with those of a classifiers if calculated from sums of per-text decisions
10	Recall$_\mu$	Multi-Class	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i+fn_i)}$	Effectiveness of a classifier to identify class labels if calculated from sums of per-text decisions
11	F-Measure$_\mu$	Multi-Class	$\frac{(\beta 2+1)\Pr ecision_\mu \operatorname{Re} call_\mu}{(\beta 2+1)\Pr ecision_\mu + \operatorname{Re} call_\mu}$	Relations between data's positive labels and those given by a classifier based on sums of per-text decisions

Deep learning method was employed to enhance classification performance in websites and text analysis [118]. Hierarchical classification in email classification provides significant steps toward improving classification performance and refining the level of security by meticulously discriminating email content.

TABLE 12. Application area wise frequency distribution of performance measures used in selected studies.

S. No.	Application Area	PRC	RCL	FMR	ACC	AUC	FPR	FNR	CTM	ERR
1	Spam	22	23	18	41	7	10	9	1	5
2	Phishing	3	3	4	7	2	3	3	1	1
3	Spam and Phishing	1	1	1	5	2	0	0	2	0
4	Multi Folder Categorization	10	10	9	18	0	1	1	1	1
5	Other	6	6	6	14	0	1	1	1	0

PRC = Precision
RCL = Recall
FMR = F-Measure
ACC = Accuracy
AUC = Area under Curve
ROC = Receiving Operator Curve
FPR = False Positive Rate
FNR = False Negative Rate
CTM = Classification Time
ERR = Error Rate

(6) Reducing processing and classification time using hardware accelerator technology: Real-time and user-centric evaluation takes relatively long processing and classification times to classify an email into particular class, which is unsuitable for real-time processing and classification [78]. Therefore, exploring the use of the hardware accelerator technology to improve processing and classification time is an interesting research direction.

(7) Dealing with the phenomenon of concept drift: Data distribution in real-time environments can change over time, thus resulting in the phenomenon of concept drift [119], [120]. A typical example of concept drift is the change in a user's interests when following an online news stream, in which the distribution of incoming news documents often remains the same. However, the conditional distribution of interesting (and not interesting) news documents for that user changes. Therefore, adaptive or incremental learning is required to update predictive models in real time to deal with concept drift. According to the current review, most studies provided solutions to email classification (for spam, phishing, and multi-folder categorization) using email content. However, email content varies with new concepts or social events. Therefore, several spam or phishing classifiers initially performed effectively, but their performances deteriorated with time. A learning email classifier is required to adjust the classification parameters for new and old emails. The problem of concept drift was addressed in classifying spam emails using the Bayesian algorithm and the incremental forgetting weight algorithm [22]. However, introducing new emails affects the classification framework. Moreover, noise may influence the framework, and the classifier can easily make incorrect decisions for cases with low numbers, particularly when several cases in a class have similar contents. Mechanisms must be introduced to address concept drift by managing noise and duplicate cases to improve classifier accuracy and performance.

(8) Reducing the false positive rate: An evaluation process may result in a false positive, which is an error indicating that a condition tested for is erroneously detected. For example, a false positive in spam email classification is a legitimate email that is mistakenly marked as spam email. The emails marked correctly or incorrectly as spam may be sent back to the sender as a bounce-email by either a server or client-side spam filters if they refuse to accept spam. The risk of accidentally marking an important legitimate email as spam may prevent companies from implementing anti-spam measures. Therefore, researchers can focus on legitimate emails misclassified as spam or phishing, so that users do not have to lose legitimate emails. A zero-defect classifier is required to avoid misclassifying legitimate emails into spam or phishing emails. The multi-stage phishing classifier using Naive Bayes, random forest, and decision trees was applied to reduce the false positive and false negative rates and improve the overall accuracy of a phishing classifier [78]. A false positive and a false negative rate of 0.4% were achieved. Nevertheless, researchers can still improve the false positive and false negative rates.

(9) Image- and text-based classification: The current review indicates that most emails are classified using text analytics. Most spammers send spam email as images. A text is inserted in an image and sent as bulk email. Therefore, spam email may be undetected. Only two out of the 98 studies [102], [103] considered images for spam email classification. In these studies, OCR-based techniques were used to convert an image into text, and 87% and 79% accuracy were achieved, respectively. OCR-based detection has some

disadvantages. The recognition is not always guaranteed to be perfect and is limited to certain fonts only. Moreover, it cannot predict CAPTCHA images and is expensive. Therefore, useful image-based features can be provided to significantly improve the performance of an email classifier.

(10) Language-based barriers: As previously discussed, five application areas were identified in the email classification domain: spam, phishing, spam and phishing, multi-folder categorization, and others. Significant work has been was conducted for spam email and phishing email classification. Researchers developed binary classifiers to categorize emails into spam or ham or into phishing or legitimate. Moreover, a ternary classifier was developed to categorize an email as spam or phishing or ham. However, the classifiers in the studies can classify emails written in English only. Only one study [98] developed a classifier that could categorize emails into spam or ham written in Chinese using the decision tree algorithm. In addition, 3 out of 98 [10], [58], [96] studies proposed classifiers that could identify the language of an email (e.g., Urdu, Hindi, and Chinese) and classify the language-specific email into a folder. None of the studies presented a phishing classifier that could categorize emails in languages other than English. Therefore, the features must be identified and developed, and a classifier that categorizes spam or phishing email written in languages other than English must be proposed.

(11) Dataset barriers and biases: Various public datasets are available for researchers on spam email classification. However, only two public datasets are accessible for phishing email classification, namely, phishing corpus and phishery corpus. Phishing corpus is used in various studies that utilize nearly 5,000 phishing emails. However, bias may result because of the low number of emails and building classifiers using one dataset. One study [80] on the phishing email classification used a custom dataset of 1,028 phishing emails. However, the number of phishing emails is too small to train the classifier for future predictions. Therefore, unlike spam email datasets, the available datasets for phishing email classification are insufficient. Making good datasets, such as the Enron dataset, publicly available can contribute to phishing email classification. In addition, all public datasets are available in English. Therefore, language bias clearly exists. Providing datasets in languages other than English can contribute to email classification. Moreover, researchers from other application areas (e.g., classification of suspicious terrorist emails) develop their own datasets of suspicious terrorist emails. As bias clearly exists in the dataset, having a standard dataset in this domain is a valuable contribution.

V. CONCLUSION

This comprehensive study presents a holistic analysis of the entire email classification domain by assembling almost all major research efforts in this regard to assist researchers in this field to gain a better understanding of the existing solutions in the major areas of email classification. Articles on email classification published in 2006–2016 were comprehensively reviewed. The selected articles were examined from five rationale aspects: email classification application areas, datasets used in each application area, features sets used in each application area, classification techniques, and performance metrics. Ninety-eight articles were rigorously selected and reviewed. Five major application areas of email classification, namely, spam, phishing, spam and phishing, multi-folder categorization, and other related application areas, were analytically summarized. A quantitative analysis of various datasets, features sets, email classification techniques, and performance measures was conducted in the identified five application areas. The most widely used datasets in the application area of spam, phishing, and multi-folder categorization were "PU," "PhishingCorpus," and "Enron," respectively. The quantitative analysis showed that the most extensively used features sets in email classification were email header part, email body part, behavioral, SpamAssasin, email URL, email JavaScript, and term-based features. In this review, five different email classification techniques were identified: supervised machine learning, semi-supervised machine learning, unsupervised machine learning, content-based learning, and statistical learning. The most widely used email classification technique was supervised machine learning technique. In the supervised machine learning technique, SVM was the most frequently used technique and showed the best performance, followed by decision trees and the Naive Bayes technique. The quantitative analysis of performance measures showed that precision, recall, accuracy, f-measure, false positive rate, false negative rate, and error rate were the frequently used measures to gauge the performance of email classifiers. Finally, 10 open research challenges for future researchers were presented.

This study has two major limitations. First, this review only focuses on email classification techniques, dataset analysis, features set analysis, and performance measure analysis. Other significant aspects, such as feature selection algorithms, feature representation techniques, feature reduction techniques, performance evaluation, and email classification tools, were not examined because of the limited scope of research. Second, the selected and reviewed articles were published from January 2006 to January 2016. The articles published after this period, if any, were not considered because of the limitation of reporting time. The scope can be extended in future reviews.

REFERENCES

[1] R. Team, "Email statistics report, 2015-2019," The Radicati Group, Inc. Palo Alto, CA, USA, Mar. 2015.

[2] J. D. Brutlag and C. Meek, "Challenges of the email domain for text classification," in *Proc. ICML*, 2000, pp. 103–110.

[3] W. W. Cohen, "Learning rules that classify e-mail," in *Proc. AAAI Spring Symp. Mach. Learn. Inf. Access*, 1996, p. 25.

[4] E. Blanzieri and 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 and 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, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proc. Anti-Phishing Work Groups 2nd Annu. Ecrime Res. Summit*, 2007, pp. 60–69.

[7] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 2070–2090, 4th Quart. 2013.

[8] Y. W. Wang, Y. N. Liu, L. Z. Feng, and X. D. Zhu, "Novel feature selection method based on harmony search for email classification," *Knowl.-Based Syst.*, vol. 73, pp. 311–323, Jan. 2015.

[9] M. R. Schmid, F. Iqbal, and B. C. M. Fung, "E-mail authorship attribution using customized associative classification," *Digit. Investigat.*, vol. 14, pp. S116–S126, Aug. 2015.

[10] M. T. Banday and S. A. Sheikh, "Multilingual e-mail classification using Bayesian filtering and language translation," in *Proc. Int. Conf. Contemp. Comput. Informat.*, 2015, pp. 696–701.

[11] M. Mohamad and A. Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," in *Proc. 2nd Int. Conf. Comput., Commun., Control Technol.*, 2015, pp. 227–231.

[12] N. A. Novino, K. A. Sohn, and T. S. Chung, "A graph model based author attribution technique for single-class e-mail classification," in *Proc. 14th IEEE/ACIS Int. Conf. Comput. Inf. Sci. (ICIS)*, Sep. 2015, pp. 191–196.

[13] W. Li, W. Meng, Z. Tan, and Y. Xiang, "Towards designing an email classification system using multi-view based semi-supervised learning," in *Proc. 13th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Sep. 2015, pp. 174–181.

[14] W. Li and W. Meng, "An empirical study on email classification using supervised machine learning in real environments," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 7438–7443.

[15] Z. J. Wang, Y. Liu, and Z. J. Wang, "E-mail filtration and classification based on variable weights of the Bayesian algorithm," in *Applied Science, Materials Science and Information Technologies in Industry*, vols. 513–517, D. L. Liu, X. B. Zhu, K. L. Xu, and D. M. Fang, Eds. Zürich, Switzerland: Trans Tech Publications Ltd., 2014, pp. 2111–2114.

[16] S. A. Saab, N. Mitri, and M. Awad, "Ham or spam? A comparative study for some content-based classification algorithms for email filtering," in *Proc. (MELECON)*, 2014, pp. 439–443.

[17] D. K. Renuka and P. Visalakshi, "Latent semantic indexing based SVM model for email spam classification," *J. Sci. Ind. Res.*, vol. 73, pp. 437–442, Jul. 2014.

[18] S. Youn, "SPONGY (SPam ONtoloGY): Email classification using two-level dynamic ontology," *Sci. World J.*, vol. 2014, pp. 1–11, Sep. 2014.

[19] Y. Meng, W. Li, and L. F. Kwok, "Enhancing email classification using data reduction and disagreement-based semi-supervised learning," in *Proc. 1st IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, USA, Apr. 2014, pp. 622–627.

[20] N. O. F. Elssied, O. Ibrahim, and W. Abu-Ulbeh, "An improved of spam e-mail classification mechanism using K-means clustering," *J. Theor. Appl. Inf. Technol.*, vol. 60, pp. 568–580, Apr. 2014.

[21] M. H. Song, "E-mail classification based learning algorithm using support vector machine," in *Materials, Mechanical Engineering and Manufacture*, vols. 268–270, H. Liu, Y. Yang, S. Shen, Z. Zhong, L. Zheng, and P. Feng, Eds. Stäfa, Switzerland: Trans Tech Publications Ltd, 2013, pp. 1844–1848.

[22] C. Jou, "Spam e-mail classification based on the IFWB algorithm," in *Proc. Intell. Inf. Database Syst.*, vol. 7802, A. Selamat, N. T. Nguyen, and H. Haron, Eds. Berlin, Germany: Springer-Verlag, 2013, pp. 314–324.

[23] T. Ma and H. Xu, "The research on email classification based on q-Gaussian kernel SVM," *J. Theor. Appl. Inf. Technol.*, vol. 48, pp. 1292–1299, Sep. 2013.

[24] J. R. Mendez, M. Reboiro-Jato, F. Diaz, E. Diaz, and F. Fdez-Riverola, "Grindstone4Spam: An optimization toolkit for boosting e-mail classification," *J. Syst. Softw.*, vol. 85, pp. 2909–2920, Dec. 2012.

[25] A. Borg and N. Lavesson, "E-mail classification using social network information," in *Proc. 7th Int. Conf. Availability, Rel. Secur. (Ares)*, 2012, pp. 168–173.

[26] N. Perez-Diaz, D. Ruano-Ordas, J. R. Mendez, J. F. Galvez, and F. Fdez-Riverola, "Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification," *Appl. Soft Comput.*, vol. 12, pp. 3671–3682, Nov. 2012.

[27] T. F. Shi, "Research on the application of e-mail classification based on support vector machine," in *Frontiers in Computer Education*, vol. 133, S. Sambath and E. Zhu, Eds. Berlin, Germany: Springer-Verlag, 2012, pp. 987–994.

[28] W. Yang and L. Kwok, "Comparison study of email classifications for healthcare organizations," in *Proc. Int. Conf. Inf. Manage. Innov. Manage. Ind. Eng. (ICIII)*, Sanya, China, 2012, pp. 468–473.

[29] L. Shi, Q. Wang, X. Ma, M. Weng, and H. Qiao, "Spam email classification using decision tree ensemble," *J. Comput. Inf. Syst.*, vol. 8, pp. 949–956, Sep. 2012.

[30] T. S. Moh and N. Lee, "Reducing classification times for email spam using incremental multiple instance classifiers," in *Proc. Inf. Intell. Syst. Technol. Manage.*, vol. 141, S. Dua, S. Sahni, and D. P. Goyal, Eds., Berlin, Germany: Springer-Verlag, 2011, pp. 189–197.

[31] V. H. Bhat, V. R. Malkani, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, "Classification of email using BeaKS: Behavior and keyword stemming," in *Proc. IEEE Region 10th Conf.*, Nov. 2011, pp. 1139–1143.

[32] J. M. Carmona-Cejudo, M. Baena-GarcÃa, J. D. Campo-Avila, and R. Morales-Bueno, "Feature extraction for multi-label learning in the domain of email classification," in *Proc. IEEE Symp. Comput. Intell. Data Mining Symp. Ser. Comput. Intell. (SSCI CIDM)*, Paris, France, Sep. 2011, pp. 30–36.

[33] R. Islam and Y. Xiang, *Email Classification Using Data Reduction Method*. New York, NY, USA: IEEE Press, 2010.

[34] J. M. Carmona-Cejudo, M. Baena-Garcia, J. del Campo-Avila, R. Morales-Bueno, and A. Bifet, "GNUsmail: Open framework for on-line email classification," in *Proc. 19th Eur. Conf. Artif. Intell.*, 2010, pp. 1141–1142.

[35] M. R. Islam, W. L. Zhou, M. Y. Guo, and X. Yang, "An innovative analyser for multi-classifier e-mail classification based on grey list analysis," *J. Netw. Comput. Appl.*, vol. 32, pp. 357–366, Mar. 2009.

[36] E.-Sayed and M. El.-Alfy, *Discovering Classification Rules for Email Spam Filtering With an ant Colony Optimization Algorithm*. New York, NY, USA: IEEE Press, 2009.

[37] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, "Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification," *Comput. Netw.*, vol. 53, pp. 835–848, Apr. 2009.

[38] J. R. Mendez, D. Glez-Pena, F. Fdez-Riverola, F. Diaz, and J. M. Corchado, "Managing irrelevant knowledge in CBR models for unsolicited e-mail classification," *Expert Syst. Appl.*, vol. 36, pp. 1601–1614, Mar. 2009.

[39] J. J. Qing, R. L. Mao, R. F. Bie, and X. Z. Gao, "An AIS-based e-mail classification method," in *Emerging Intelligent Computing Technology and Applications: With Aspects of Artificial Intelligence*, vol. 5755, D. S. Huang, K. H. Jo, H. H. Lee, V. Bevilacqua, and H. J. Kang, Eds. Berlin, Germany: Springer-Verlag, 2009, pp. 492–499.

[40] B. Yu and D. H. Zhu, "Combining neural networks and semantic feature space for email classification," *Knowl.-Based Syst.*, vol. 22, pp. 376–381, Sep. 2009.

[41] Y. F. Yi, C. H. Li, and W. Song, *Email Classification Using Semantic Feature Space*. Los Alamitos, CA, USA: IEEE Computer Soc, 2008.

[42] R. Islam, W. L. Zhou, and M. U. Chowdhury, *Email Categorization Using (2+1)-Tier Classification Algorithms*. Los Alamitos, CA, USA: IEEE Computer Soc, 2008.

[43] M. R. Islam, J. Singh, A. Chonka, and W. Zhou, *Multi-Classifier Classification of Spam Email on an Ubiquitous Multi-Core Architecture*. Los Alamitos, CA, USA: IEEE Computer Soc, 2008.

[44] Z. Q. Zhu, *An Email Classification Model Based on Rough Set and Support Vector Machine*. Los Alamitos, CA, USA: IEEE Computer Soc, 2008.

[45] M. Balakumar and V. Vaidehi, *Ontology Based Classification and Categorization of Email*. New York, NY, USA: IEEE Press, 2008.

[46] C. C. Lai and C. H. Wu, "Particle swarm optimization-aided feature selection for spam email classification," in *Proc. 2nd Int. Conf. Innov. Comput. Inf. Control, (ICICIC)*, Kumamoto, Japan, 2007, p. 165.

[47] K. Yelupula and S. Ramaswamy, "Social network analysis for email classification," in *Proc. 46th Annu. Southeast Regional Conf. (ACM-SE)*, Auburn, AL, USA, 2008, pp. 469–474.

[48] S. Youn and D. McLeod, "Spam email classification using an adaptive ontology," *J. Softw.*, vol. 2, pp. 43–55, Sep. 2007.

[49] T. L. Wong, K. O. Chow, and F. Wong, *Incorporting Keyword-Based Filtering to Document Classification for Email Spamming*. New York, NY, USA: IEEE Press, 2007.

[50] M. R. I. Wanlei and W. L. Zhou, *Email Categorization Using Multi-Stage Classification Technique*. Los Alamitos, CA, USA: IEEE Computer Soc, 2007.

[51] T. Ichimura, A. Hara, Y. Kurosawa, T. Ichimura, A. Hara, and Y. Kurosawa, "A classification method for spam e-mail by self-organizing map and automatically defined groups," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vols. 1–8. Oct. 2007, pp. 310–315.

[52] S. Youn and D. McLeod, *A Comparative Study for Email Classification*. Dordrecht, The Netherlands: Springer, 2007.

[53] S. Misina, "Incremental learning for e-mail classification," in *Computational Intelligence, Theory and Application*, B. Ruesch, Ed. Berlin, Germany: Springer-Verlag, 2006, pp. 545–553.

[54] M. N. Marsono, M. W. El-Khaxashi, F. Gebali, and S. Ganti, *Distributed Layer-3 E-mail Classification for SPAM Control*. New York, NY, USA: IEEE Press, 2006.

[55] M. R. Islam and W. L. Zhou, "Minimizing the limitations of GL analyser of fusion based email classification," in *Algorithms and Architectures for Parallel Processing, Proceedings*, vol. 5574, A. Hua and S. L. Chang, Eds. Berlin, Germany: Springer-Verlag, 2009, pp. 761–774.

[56] M. R. Islam, J. Abawajy, and M. Warren, *Multi-Tier Phishing Email Classification with an Impact of Classifier Rescheduling*. New York, NY, USA: IEEE, 2009.

[57] K. Xu, C. Wen, Q. Yuan, X. He, and J. Tie, "A mapreduce based parallel SVM for email classification," *J. Netw.*, vol. 9, pp. 1640–1647, Sep. 2014.

[58] M. T. Banday and S. A. Sheikh, "Folder classification of urdu and hindi language e-mail messages," in *Proc. 3rd Int. Conf. Comput. Knowl. Eng. (Iccke)*, 2013, pp. 59–63.

[59] M. Li, Y. Park, R. Ma, and H. Y. Huang, "Business email classification using incremental subspace learning," in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Tsukuba, Japan, 2012, pp. 625–628.

[60] M. F. Wang, M. F. Tsai, S. L. Jheng, and C. H. Tang, "Social feature-based enterprise email classification without examining email contents," *J. Netw. Comput. Appl.*, vol. 35, pp. 770–777, Apr. 2012.

[61] A. Bacchelli, T. Dal Sasso, and M. D'Ambros, and M. Lanza, "Content classification of development Emails," in *Proc. 34th Int. Conf. Softw. Eng.*, 2012, pp. 375–385.

[62] I. Alberts and D. Forest, "Email pragmatics and automatic classification: A study in the organizational context," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, pp. 904–922, May 2012.

[63] A. A. Al Sallab and M. A. Rashwan, "E-mail classification using deep networks," *J. Theor. Appl. Inf. Technol.*, vol. 37, pp. 241–251, Sep. 2012.

[64] J. Pujara, H. Daumé, III, and L. Getoor, "Using classifier cascades for scalable e-mail classification," in *Proc. 8th Annu. Collaboration, Electron. Messaging, Anti–Abuse Spam Conf. (CEAS)*, Perth, WA, USA, Sep. 2011, pp. 55–63.

[65] N. Chatterjee, S. Kaushik, S. Rastogi, and V. Dua, "Automatic email classification using user preference ontology," in *Proc. Int. Conf. Knowl. Eng. Ontol. Develop. (KEOD)*, Valencia, Spain, Sep. 2010, pp. 165–170.

[66] D. M. Jones, *Learning to Improve E-mail Classification With Numero Interactive*. London, U.K.: Springer-Verlag, 2010.

[67] S. Chakravarthy, A. Venkatachalam, and A. Telang, "A graph-based approach for multi-folder email classification," in *Proc. 10th IEEE Int. Conf. Data Mining, (ICDM)*, Sydney, NSW, USA, Sep. 2010, pp. 78–87.

[68] D. M. Jones, "Learning to improve e-mail classification with numáro interactive," in *Proc. 29th SGAI Int. Conf. Innov. Techn. Appl. Artif. Intell.*, Cambridge, MA, USA, 2010, pp. 377–390.

[69] S. Park and D. U. An, "Automatic e-mail classification using dynamic category hierarchy and semantic features," *IETE Techn. Rev.*, vol. 27, pp. 478–492, Apr. 2010.

[70] S. Baskaran, "Content based email classification system by applying conceptual maps," in *Proc. Int. Conf. Intell. Agent Multi–Agent Syst. (IAMA)*, 2009, pp. 1–2.

[71] M. Chang and C. K. Poon, "Using phrases as features in email classification," *J. Syst. Softw.*, vol. 82, pp. 1036–1045, Sep. 2009.

[72] A. Krzywicki and W. Wobcke, "Incremental e-mail classification and rule suggestion using simple term statistics," in *Advances in Artificial Intelligence*, vol. 5866, A. Nicholson and X. Li, Eds. Berlin, Germany: Springer-Verlag, 2009, pp. 250–259.

[73] T. Ayodele, S. Zhou, and R. Khusainov, "Email classification: Solution with back propagation technique," in *Proc. Int. Conf. Int. Technol. Secured Trans. (ICITST)*, London, U.K., 2009, pp. 1–6.

[74] C. Zeng, J. Gu, and Z. Lu, "A new approach to email classification using concept vector space model," in *Proc. 2nd Int. Conf. Future Generat. Commun. Netw. Symp. (FGCN)*, 2008, pp. 162–166.

[75] T. Ayodele, R. Khusainov, and D. Ndzi, "Email classification and summarization: A machine learning approach," in *Proc. IET Conf. Wireless, Mobile Sensor Netw. (CCWMSN)*, Shanghai, China, 2007, pp. 805–808.

[76] J. Stefanowski and M. Zienkowicz, "Classification of polish email messages: Experiments with various data representations," in *Foundations of Intelligent Systems, Proceedings*, vol. 4203, F. Esposito, Z. W. Ras, D. Malerba, and G. Semeraro, Eds. Berlin, Germany: Springer-Verlag, 2006, pp. 723–728.

[77] A. A. Akinyelu and A. O. Adewumi, "Classification of phishing email using random forest machine learning technique," *J. Appl. Math.*, vol. 2014, pp. 1–6, Apr. 2014.

[78] A. Y. Daeef, R. B. Ahmad, Y. Yacob, N. Yaakob, and K. N. F. K. Azir, "Multi stage phishing email classification," *J. Theor. Appl. Inf. Technol.*, vol. 83, pp. 206–214, Sep. 2016.

[79] R. Dazeley, J. L. Yearwood, B. H. Kang, and A. V. Kelarev, "Consensus clustering and supervised classification for profiling phishing emails in Internet commerce security," in *Knowledge Management and Acquisition for Smart Systems and Services*, vol. 6232, B. H. Kang and D. Richards, Eds. Berlin, Germany: Springer-Verlag, 2010, pp. 235–246.

[80] M. D. del Castillo, A. Iglesias, and J. I. Serrano, "Detecting phishing e-mails by heterogeneous classification," in *Intelligent Data Engineering and Automated Learning—Ideal*, vol. 4881, H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 296–305.

[81] I. R. A. Hamid, J. Abawajy, and T. H. Kim, "Using feature selection and classification scheme for automating phishing email detection," *Stud. Informat. Control*, vol. 22, pp. 61–70, Mar. 2013.

[82] M. Khonji, A. Jones, and Y. Iraqi, "An empirical evaluation for feature selection methods in phishing email classification," *Comput. Syst. Sci. Eng.*, vol. 28, pp. 37–51, Apr. 2013.

[83] I. Qabajeh and F. Thabtah, "An experimental study for assessing email classification attributes using feature selection methods," in *Proc. 3rd Int. Conf. Adv. Comput. Sci. Appl. Technol. (ACSAT)*, 2014, pp. 125–132.

[84] M. Zareapoor, P. Shamsolmoali, and M. A. Alam, "Highly discriminative features for phishing email classification by SVD," in *Information Systems Design and Intelligent Applications*, vol. 339, J. K. Mandal, S. C. Satapathy, M. K. Sanyal, P. P. Sarkar, and A. Mukhopadhyay, Eds. Berlin, Germany: Springer-Verlag, 2015, pp. 649–656.

[85] S. Smadi, N. Aslam, L. Zhang, R. Alasem, and M. A. Hossain, "Detection of phishing emails using data mining algorithms," in *Proc. 9th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (Skima)*, 2015, pp. 1–8.

[86] J. C. Gomez and M. F. Moens, "PCA document reconstruction for email classification," *Comput. Statist. Data Anal.*, vol. 56, pp. 741–751, Sep. 2012.

[87] J. C. Gomez, E. Boiy, and M.-F. Moens, "Highly discriminative statistical features for email classification," *Knowl. Inf. Syst.*, vol. 31, pp. 23–53, Apr. 2012.

[88] A. G. K. Janecek and W. N. Gansterer, "E-mail classification based on NMF," in *Proc. 9th SIAM Int. Conf. Data Mining (SDM)*, Sparks, NV, USA, 2009, pp. 1345–1354.

[89] W. N. Gansterer and D. Polz, "E-mail classification for phishing defense," in *Advances in Information Retrieval*, vol. 5478, M. Boughanem, C. Berrut, J. Mothe, and C. SouleDupuy, Eds. Berlin, Germany: Springer-Verlag, 2009, pp. 449–460.

[90] W. Zhao and Y. Zhu, "An email classification scheme based on decision-theoretic rough set theory and analysis of email security," in *Proc. IEEE Region 10th Conf. (TENCON)*, 2007, pp. 1–6.

[91] H. Dalianis, J. Sjöbergh, and E. Sneiders, "Comparing manual text patterns and machine learning for classification of e-mails for automatic answering by a government agency," in *Computational Linguistics and Intelligent Text Processing*, vol. 6609, A. Gelbukh, Ed. Berlin, Germany: Springer-Verlag, 2011, pp. 234–243.

[92] K. Iwai, K. Iida, M. Akiyoshi, and N. Komoda, "A classification method of inquiry e-mails for describing FAQ with self-configured class dictionary," in *Distributed Computing and Artificial Intelligence*, vol. 79, A. P. D. DeCarvalho, S. RidriguezGonzalez, J. F. D. Santana, and J. M. C. Rodriguez, Eds. Berlin, Germany: Springer-Verlag, 2010, pp. 35–43.

[93] R. Tailby, R. Dean, B. Milner, and D. Smith, "Email classification for automated service handling," in *Proc. ACM Symp. Appl. Comput.*, Dijon, France, 2006, pp. 1073–1077.

[94] K. Karthik and R. Ponnusamy, "Adaptive machine learning approach for emotional email classification," in *Human-Computer Interaction: Towards Mobile and Intelligent Interaction Environments*, vol. 6763, J. A. Jacko, Ed. Berlin, Germany: Springer-Verlag, 2011, pp. 552–558.

[95] K. Coussement and D. Van den Poel, "Improving customer complaint management by automatic email classification using linguistic style features as predictors," *Decision Support Syst.*, vol. 44, pp. 870–882, Sep. 2008.

[96] M. T. Banday and S. A. Sheikh, "Realization of microsoft outlook (R) add-in for language based e-mail folder classification," in *Proc. Int. Conf. Mach. Intell. Res. Adva. (Icmira)*, 2013, pp. 279–284.

[97] N. Al Fe'ar, E. Al Turki, A. Al Zaid, M. Al Duwais, M. Al Sheddi, and N. Al Khamees, *E-Classifier: A Bi-Lingual Email Classification System*. New York, NY, USA: IEEE, 2008.

[98] H. Chen, Y. Zhan, and Y. Li, "The application of decision tree in Chinese email classification," in *Proc. Int. Conf. Mach. Learn. (ICMLC)*, Qingdao, China, 2010, pp. 305–308.

[99] E. K. Jamison and I. Gurevych, "Headerless, quoteless, but not hopeless? Using pairwise email classification to disentangle email threads," in *Proc. 9th Int. Conf. Recent Adv. Natural Lang. Process.*, Hissar, India, 2013, pp. 327–335.

[100] N. Prattipati and E. Hart, "Evaluation and extension of the AISEC email classification system," in *Artificial Immune Systems*, vol. 5132, P. J. Bentley, D. Lee, and S. Jung, Eds. Berlin, Germany: Springer-Verlag, 2008, pp. 154–165.

[101] M. F. Wang, S. L. Jheng, M. F. Tsai, and C. H. Tang, "Enterprise email classification based on social network features," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, 2011, pp. 532–536.

[102] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, "Text and image based spam email classification using KNN, naive Bayes and reverse DBSCAN algorithm," in *Proc. Int. Conf. Reliabilty, Optim., Inf. Technol. (Icroit)*, 2014, pp. 153–155.

[103] J. H. Hsia and M. S. Chen, "Language-model-based detection cascade for efficient classification of image-based SPAM e-mail," in *Proc. IEEE Int. Conf. Multimedia Expo*, vols. 1–3. Sep. 2009, pp. 1182–1185.

[104] S. Appavu, R. Rajaram, M. Muthupandian, G. Athiappan, and K. S. Kashmeera, "Data mining based intelligent analysis of threatening e-mail," *Knowl.-Based Syst.*, vol. 22, pp. 392–393, Jul. 2009.

[105] P. Zhang, J. L. Zhang, and Y. Shi, "A new multi-criteria quadratic-programming linear classification model for VIP e-mail analysis," in *Computational Science—ICCS*, vol. 4488, Y. Shi, G. D. VanAlbada, J. Dongarra, and P. M. A. Sloot, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 499–502.

[106] G. V. Cormack, "Email spam filtering: A systematic review," *Found. Trends Inf. Retr.*, vol. 1, pp. 335–455, Sep. 2007.

[107] J. Nazario, "PhishingCorpus homepage," *Retrieved*, Apr. 2006. [Online]. Available: http://monkey.org/%7Ejose/wiki/doku.php?id=PhishingCorpus

[108] H. Gonzalez, K. Nance, and J. Nazario, "Phishing by form: The abuse of form sites," in *Proc. 6th Int. Conf. Malicious Unwanted Softw. (MALWARE)*, 2011, pp. 95–101.

[109] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *Proc. Eur. Conf. Mach. Learn.*, vol. 3201, 2004, pp. 217–226.

[110] M. A. Oveis-Gharan and K. Raahemifar, "Multiple classifications for detecting spam email by novel consultation algorithm," in *Proc. IEEE 27th Can. Conf. Elect. Comput. Eng.*, New York, NY, USA, 2014, pp. 1–5.

[111] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine Learning: An Artificial Intelligence Approach*. New York, NY, USA: Springer, 2013.

[112] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.*, 1998, pp. 137–142.

[113] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[114] F. J. Provost and T. Fawcett, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions," in *Proc. KDD*, 1997, pp. 43–48.

[115] F. J. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proc. ICML*, 1998, pp. 445–453.

[116] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[117] A. Zhang, L. G. Pueyo, J. B. Wendt, M. Najork, and A. Broder, *Email Category Prediction*. New York, NY, USA: Association for Computing Machinery (ACM), 2017.

[118] S. Dumais and H. Chen, "Hierarchical classification of Web content," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2000, pp. 256–263.

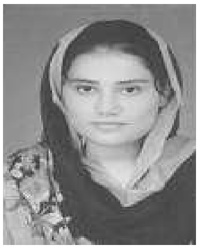
[119] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Mach. Learn.*, vol. 23, pp. 69–101, Sep. 1996.

[120] J. C. Schlimmer and R. H. Granger, Jr., "Incremental learning from noisy data," *Mach. Learn.*, vol. 1, pp. 317–354, Oct. 1986.

GHULAM MUJTABA received the master's degree (Hons.) in computer science from National University, FAST, Karachi, Pakistan. He is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has been an Assistant Professor with the Sukkur Institute of Business Administration (Sukkur IBA), Sukkur, Pakistan, since 2006. He has vast experience in teaching and research. Prior Sukkur IBA, he was with well-known software house in Karachi, Pakistan, for four years. He has authored or co-authored several articles in academic journals indexed in well reputed databases, such as ISI-indexed and Scopus-indexed. His field of research is machine learning, online social networking, text mining, deep learning, and information retrieval.

LIYANA SHUIB received the Ph.D. degree in computer science from the University of Malaya, Malaysia, in 2010. She has over eight years of experience as an Academician and also successfully supervised Ph.D. and master's students. She currently supervises eight master's and five Ph.D. students and has examined master's and Ph.D. theses. She is currently a Senior Lecturer with the Department of Information Systems, University of Malaya. She has authored or co-authored several articles in academic journals indexed in well reputed databases, such as ISI-indexed and Scopus-indexed. Her research area includes information retrieval and decision support in information systems.

RAM GOPAL RAJ received the Ph.D. degree in computer science from the University of Malaya, Malaysia, in 2010. He has over ten years of experience as an Academician and also successfully supervised Ph.D. and master's students. He currently supervises five master's and six Ph.D. students and has examined master's and Ph.D. theses. He is currently a Senior Lecturer with the Department of Artificial intelligence, University of Malaya. He has authored or co-authored 30 ISI-indexed research articles. His research area includes text classification, natural language processing, logic programming, and intelligent systems.



NAHDIA MAJEED received the master's degree (Hons.) in management information system from the Sukkur Institute of Business Administration (Sukkur IBA), Sukkur, Pakistan. She has vast experience in teaching and research. She is currently an Assistant Professor with Sukkur IBA. Her area of interest is mathematical modeling, collaborative software process measurements, and mapping data from object oriented databases to RDF-based ontology.

MOHAMMED ALI AL-GARADI received the M.Tech. degree in electronic and communication engineering from Jawaharlal Nehru Technological University, Hyderabad, India. He is currently pursuing the Ph.D. degree with the Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has authored or co-authored several articles in academic journals indexed in well reputed databases, such as ISI-indexed and Scopus-indexed. His field of research is online social networking, text mining, deep learning, and information retrieval.