DISCOVERY OF SPAM IN COMMMUNAL NETWORK

S.SAHARIKA., M.E(CSE)., VEERAMMAL COLLEGE OF ENGINEERING, SINGARAKOTAI,DINDUGAL.

2ND AUTHOR:K.SRIDAR., M.E.,(Ph.D)., ASSISTANT PROFESSOR ,CSE,VEERAMMAL COLLEGE OF ENGINEERING,SINGARAKOTAI,DINDUGAL.

3RD AUTHOR:S.AURTHY FELLICITA.,M.E.,

ASSISTANT PROFESSOR,VEERAMMAL COLLEGE OF ENGINEERING, SINGARAKOTAI,DINDUGAL.

Abstract—

Personal and business users prefer to use email as one of the crucial sources of communication. The usage and importance of emails continuously grow despite the prevalence of alternative means, such as electronic messages, mobile applications, and social networks. As the volume of business-critical emails continues to grow, the need to automate the management of emails increases for several reasons, such as spam email classification, phishing email classification, and multi-folder categorization, among others. This study comprehensively reviews articles on email classification published in 2006–2016 by exploiting the methodological decision analysis in five aspects, namely, email classification application areas, datasets used in each application area, feature space utilized in each application area, email classification techniques, and 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 email classification was applied. Moreover, various public datasets, features sets, classification techniques, and performance measures are examined and used in each identified application area. This review identifies five application areas of email classification. The most widely used datasets, features sets, classification techniques, and performance measures are found in the identified application areas. The extensive use of these popular datasets, features sets, classification techniques, and performance measures is discussed and justified. The research directions, research challenges,and open issues in the field of email 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 isbecoming the most extensively used communicationmechanism. In recent years, the increased use of emails hasled to the emergence and further escalation of problemscaused by spam and phishing emails. A typical user receivesabout 40–50 emails per day for others, hundreds ofmessages are usual. Users spend a significant part of workingtime on processing emails. Therefore, email management is animportant issue faced by organizations and individuals, and itnecessitates the need to devise mechanisms that intelligentlyclassify and deal with the problem. Generally, the main toolfor email management is automatic email classification.An automatic email classifier is a system that automaticallyclassifies emails into one or more of a discrete set ofpredefined categories. For instance, for email management,one can benefit from a system that classifies

an incomingemail into official or personal, phishing or normal, and spamor 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. 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 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.

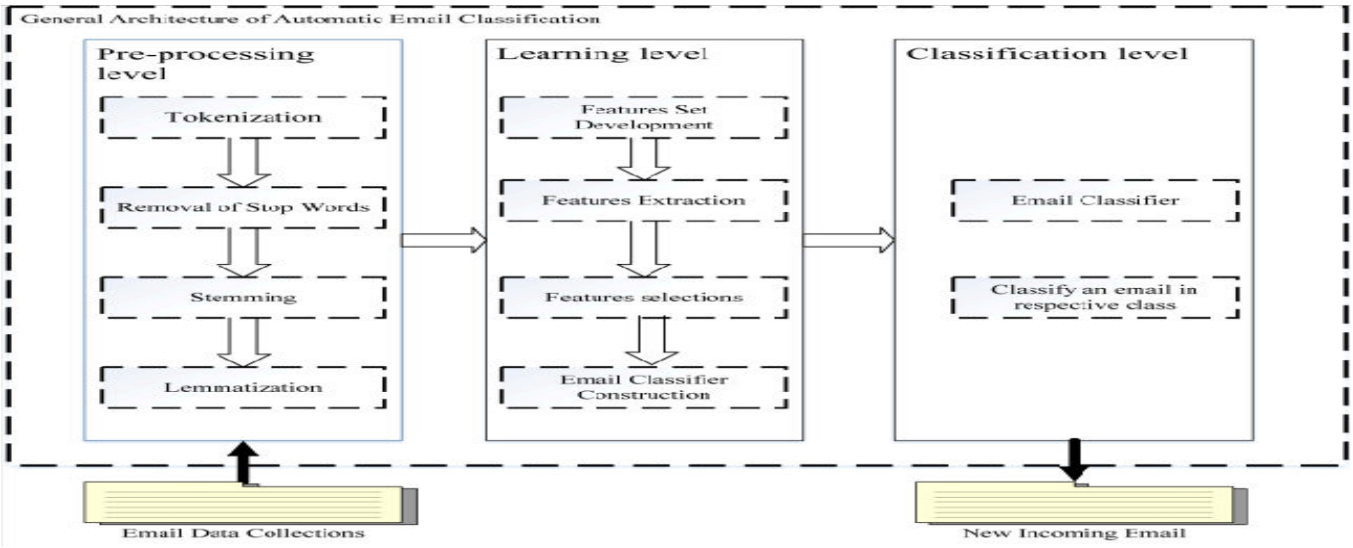

Figure 1: General architecture of automatic email classification

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. To predict phishing emails, Abu-Nimeh et al. compared the predictive accuracy of numerous machine learning algorithms, including logistic regression, classification and regression trees, supportvector machines (SVMs), random forest, and neural networks. Almomani et al. 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.

(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 beenapplied?

(2)Which publicly available datasets can be accessed for the various application areas of

emailclassification?
(3)What are the widely used features in the various application areas of emailclassification?
(4)What are the widely used machine learning techniques in the area of emailclassification?
(5)What performance evaluation metrics are employed to evaluate email classifierperformance?
(6)What are the challenges and future research directions for future researchers working in the email classification domain?

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 Scopusdatabases.
(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

(4) This review. The articles from 2006 were selected because this field became popular in that year.
(5)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.
(6) Only articles published in the English language were extracted.

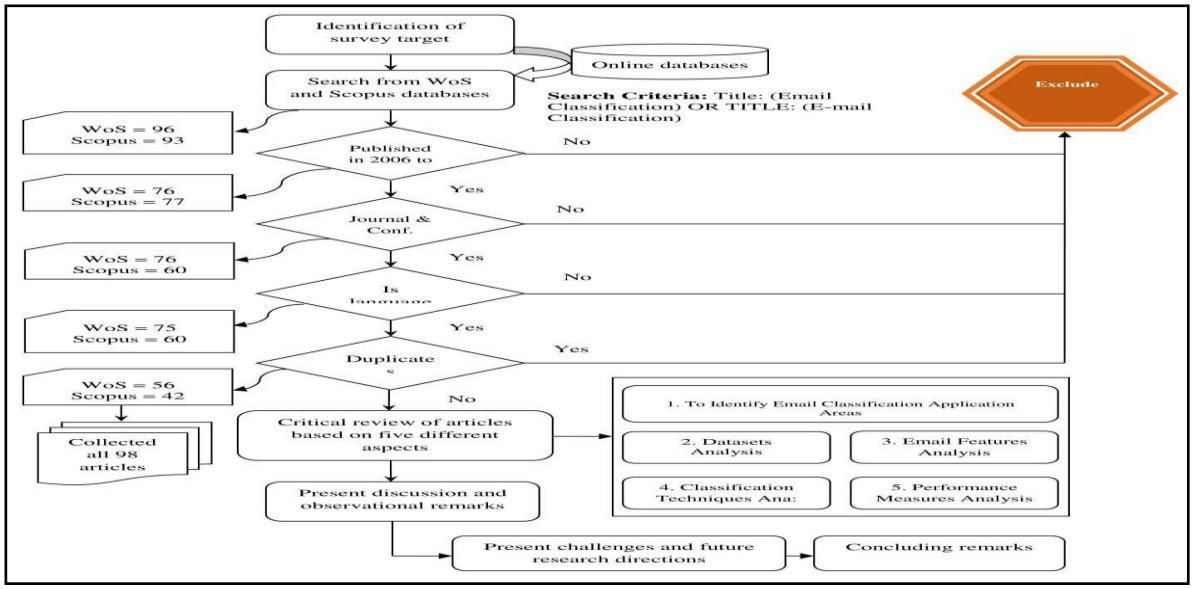

Figure 2: Research Methodology

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 76from 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 areasare 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, areincluded. 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 ornormal, 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.

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 Table2.

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]. 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.

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 in email classification are email header, email body, email JavaScript, email URL, behavioral, Spam Assasin, 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:

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.

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
	PU	10	Total 7101 email (spam = 3020 and ham = 4081)	[13, 19, 33, 35, 40-43, 50, 55]	
	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]	
Spam Email Classification	Enron Spam Corpus	5	Total 30041 emails (spam = 13496 and ham = 16545)	[9, 12, 32, 34, 47]	49
	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 SpamAssasin = 6951)	[56, 77-84]	9
	Custom	1	Total emails 2034 emails (phishing = 1028 , ham = 1006)	[80]	

Spam and Phishing Email Classification	PU, LingSpam, SpamAssassin, 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]	

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 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, amongothers.

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 orimages), 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.

Social Features:

Social features include work related and work unrelated social relationships of employees during

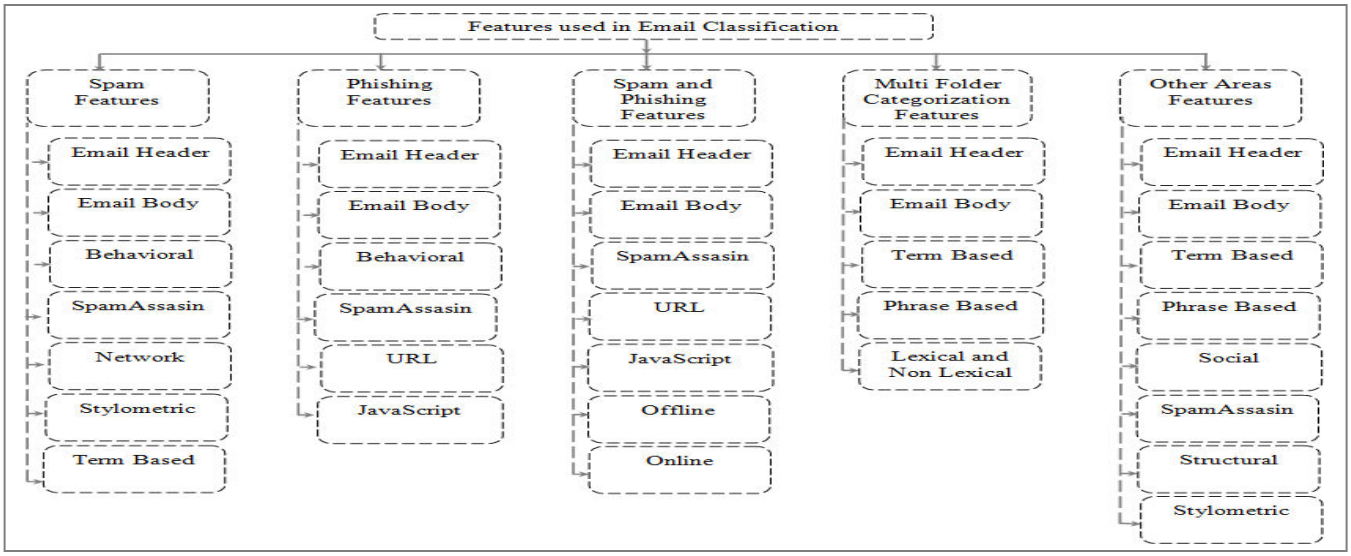

Figure 5: Feature set taxonomy used in email classification

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 torecipient social network).Whilelexicalfeaturesinclude: 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 Spam Assasin 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 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. Spam Assasin features are specially designed to detect spam emails.

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. 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 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.

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 techniqueamongallthelistedmethods.Outofthe98studies, 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 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.

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

Table 10: Confusion Matrix

IV. FUTURE RESEARCHDIRECTIONS

This section highlights several research challenges and open issues in the current studies on email classification. In this

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 underCurve
ROC = Receiving OperatorCurve FPR = False PositiveRate
FNR = False Negative Rate CTM = ClassificationTime ERR = ErrorRate

(1) 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.

(2) Deep learning: Deep learning enables computational methods with several processing layers to learn representations of data with several levels of abstraction 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 givendata.

(3) 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.

(4) 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 researchdirection.

(5) 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 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.

(6) 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.

(7) 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 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.

(8) 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.

(9) 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 on the phishing email classification used a custom dataset of 1,028 phishing emails.

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,” “Phishing Corpus,” 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 e-mail 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 views.

ACKNOWLEDGMENT

This research was partially funded by the University Malaya Research Grant (Grant No:RP026-14AET).

REFERENCES

[1] Radicati, "Email Statistics Report, 2015-2019," THE RADICATI GROUP,INC.2015.

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

[3] W. W. Cohen, "Learning rules that classify e-mail," in *AAAI spring symposium on machine learning in information access*, 1996, p. 25.

[4] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," *Artificial Intelligence Review,* vol. 29, pp. 63-92,2008.

[5] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications,* vol. 36, pp. 10206-10222,2009.

[6] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," in *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, 2007, pp.60-69.

[7] A. Almomani, B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," *Communications Surveys & Tutorials, IEEE,* vol. 15, pp. 2070- 2090,2013.

[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," *Knowledge-Based Systems,* vol. 73, pp. 311-323, Jan2015.

[9] M. R. Schmid, F. Iqbal, and B. C. M. Fung, "E-mail authorship attribution using customized associative classification," *Digital Investigation,* vol. 14, pp. S116-S126, Aug2015.

[10] M. T. Banday and S. A. Sheikh, "Multilingual e-mail classification using Bayesian filtering and language translation," in *2014 International Conference on Contemporary Computing and Informatics, IC3I 2014*, 2015, pp.696-701.

[11] M. Mohamad and A. Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," in *2nd International Conference on Computer, Communications, and Control Technology, I4CT 2015*, 2015, pp.227-231.

[12] N. A. Novino, K. A. Sohn, and T. S. Chung, "A graph model based author attributiontechnique for single-class e-mail classification,"

in *14th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2015*, 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 *13th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2014*, 2015, pp.174-181.

[14] W. Li and W. Meng, "An empirical study on email classification using supervised machine learning in real environments," in *IEEE International Conference on Communications, ICC 2015*, 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*. vol. 513-517, D. L. Liu, X. B. Zhu, K. L. Xu, and D. M. Fang, Eds., ed Stafa-Zurich: Trans Tech Publications Ltd, 2014, pp.2111-2114.

[16] S. A. Saab, N. Mitri, M. Awad, and Ieee, "Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering," *2014 17th Ieee Mediterranean Electrotechnical Conference (Melecon),* pp. 439-443,2014.

[17] D. K. Renuka and P. Visalakshi, "Latent Semantic Indexing Based SVM Model for Email Spam Classification," *Journal of Scientific & Industrial Research,* vol. 73, pp. 437-442, Jul2014.

[18] S. Youn, "SPONGY (SPam ONtoloGY): Email classification using two-level dynamic ontology," *Scientific World Journal,* vol. 2014, 2014.

[19] Y. Meng, W. Li, and L. F. Kwok, "Enhancing email classification using data reduction and disagreement-based semi-supervised learning," in *2014 1st IEEE International Conference on Communications, ICC 2014*, Sydney, NSW, 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," *Journal of Theoretical and Applied Information Technology,* vol. 60, pp. 568-580,2014.

[21] M. H. Song, "E-Mail Classification based Learning Algorithm Using Support vector machine," in *Materials, Mechanical Engineering and Manufacture, Pts 1-3*. vol. 268-270, H. Liu, Y. Yang, S. Shen, Z. Zhong, L. Zheng, and P. Feng, Eds., ed Stafa-Zurich: Trans Tech Publications Ltd, 2013, pp.1844-1848.

[22] C. Jou, "Spam E-Mail Classification Based on the IFWB Algorithm," in *Intelligent Information and Database Systems*. vol. 7802, A. Selamat, N. T. Nguyen, and H. Haron, Eds., ed Berlin: Springer-Verlag Berlin, 2013, pp.314-324.

[23] Lifan, T. Ma, and H. Xu, "The research on email classification based on q-Gaussian kernel SVM," *Journal of Theoretical and Applied Information Technology,* vol. 48, pp. 1292-1299,2013.

[24] J. R. Mendez, M. Reboiro-Jato, F. Diaz, E. Diaz, and F. Fdez- Riverola, "Grindstone4Spam: An optimization toolkit for boosting e-mail classification," *Journal of Systems and Software,* vol. 85, pp. 2909-2920, Dec2012.

[25] A. Borg, N. Lavesson, and Ieee, "E-mail Classification using Social Network Information," *2012 Seventh International Conference on Availability, Reliability and Security (Ares),* pp. 168-173,2012.

[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," *Applied Soft Computing,* vol. 12, pp. 3671-3682, Nov2012.

[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., ed Berlin: Springer-Verlag Berlin, 2012, pp.987-994.

[28] W. Yang and L. Kwok, "Comparison study of email classifications for healthcare organizations," in *2012 International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2012*, Sanya, 2012, pp.468-473.

[29] L. Shi, Q. Wang, X. Ma, M. Weng, and H. Qiao, "Spam email classification using decision tree ensemble," *Journal of Computational Information Systems,* vol. 8, pp. 949-956,2012.

[30] T. S. Moh and N. Lee, "Reducing Classification Times for Email Spam Using Incremental Multiple Instance Classifiers," in *Information Intelligence, Systems, Technology and Management*. vol. 141, S. Dua, S. Sahni, and D. P. Goyal, Eds., ed Berlin: Springer-Verlag Berlin, 2011, pp.189-197.

[31] V. H. Bhat, V. R. Malkani, P. D. Shenoy, K. R. Venugopal, L. M. Patnaik, and Ieee, "Classification of Email using BeaKS: Behavior and Keyword Stemming," *2011 Ieee Region 10 Conference Tencon 2011,* pp. 1139-1143,2011.

[32] J.M.Carmona-Cejudo,M.Baena-García,J.D.Campo-Avila,and R. Morales-Bueno, "Feature extraction for multi-label learning in the domain of email classification," in *Symposium Series on Computational Intelligence, IEEE SSCI2011 - 2011 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011*, Paris, 2011, pp.30-36.

[33] R. Islam, Y. Xiang, and Ieee, *Email Classification Using Data Reduction Method*. New York: Ieee,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 *Ecai 2010 - 19th European Conference on Artificial Intelligence*. vol. 215, H. Coelho, R. Studer, and M. Wooldridge, Eds., ed Amsterdam: Ios Press, 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," *Journal of Network and Computer Applications,* vol. 32, pp. 357-366, Mar2009.

[36] E. S. M. El-Alfy and Ieee, *Discovering Classification Rules for Email Spam Filtering with an Ant Colony Optimization Algorithm*. New York: Ieee,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," *Computer Networks,* vol. 53, pp. 835- 848, Apr2009.

[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 Systems with Applications,* vol. 36, pp. 1601-1614, Mar2009.

[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., ed Berlin: Springer-Verlag Berlin, 2009, pp.492-499.

[40] B. Yu and D. h. Zhu, "Combining neural networks and semantic feature space for email classification," *Knowledge-Based Systems,* vol. 22, pp. 376-381,2009.

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

[42] R. Islam, W. L. Zhou, and M. U. Chowdhury, *Email categorization using (2+1)-tier classification algorithms*. Los Alamitos: 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: Ieee Computer Soc,2008.

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

[45] M. Balakumar, V. Vaidehi, and Ieee, *Ontology based classification and categorization of email*. New York: Ieee,2008.

[46] C. C. Lai and C. H. Wu, "Particle swarm optimization-aided feature selection for spam email classification," in *2nd International Conference on Innovative Computing, Information and Control, ICICIC 2007*, Kumamoto,2008.

[47] K. Yelupula and S. Ramaswamy, "Social network analysis for email classification," in *46th Annual Southeast Regional*

Conference on XX, ACM-SE 46, Auburn, AL, 2008, pp.469-474.

[48] S. Youn and D. McLeod, "Spam email classification using an adaptive ontology," *Journal of Software,* vol. 2, pp. 43-55,2007.

[49] T. L. Wong, K. O. Chow, and F. Wong, *Incorporting keyword- based filtering to document classification for email spamming*. New York: Ieee,2007.

[50] M. R. I. Wanlei and W. L. Zhou, *Email categorization using multi- stage classification technique*. Los Alamitos: Ieee Computer Soc, 2007.

[51] T. Ichimura, A. Hara, Y. Kurosawa, and Ieee, "A classification method for seam E-mail by Self-Organizing Map and automatically defined groups," in *2007 Ieee International Conference on Systems, Man and Cybernetics, Vols 1-8*, ed New York: Ieee, 2007, pp.310-315.

[52] S. Youn and D. McLeod, *A comparative study for email classification*. Dordrecht: Springer,2007.

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

[54] M. N. Marsono, M. W. El-Khaxashi, F. Gebali, S. Ganti, and Ieee, *Distributed layer-3 e-mail classification for SPAM control*. New York: Ieee,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., ed Berlin: Springer-Verlag Berlin, 2009, pp.761-774.

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

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

[58] M. T. Banday, S. A. Sheikh, and Ieee, "Folder Classification of Urdu and Hindi Language E-mail Messages," *Proceedings of the 3rd International Conference on Computer and Knowledge Engineering (Iccke 2013),* pp. 59-63,2013.

[59] M. Li, Y. Park, R. Ma, and H. Y. Huang, "Business email classification using incremental subspace learning," in *21st International Conference on Pattern Recognition, ICPR 2012*, Tsukuba, 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," *Journal of Network and Computer Applications,* vol. 35, pp. 770-777,2012.

[61] A. Bacchelli, T. Dal Sasso, M. D'Ambros, and M. Lanza, "Content Classification of Development Emails," in *2012 34th International ConferenceonSoftwareEngineering*,M.Glinz,G.Murphy,and M. Pezze, Eds., ed New York: Ieee, 2012, pp.375-385.

[62] I. Alberts and D. Forest, "Email pragmatics and automatic classification: A study in the organizational context," *Journal of the American Society for Information Science and Technology,* vol. 63, pp. 904-922, May2012.

[63] A. A. Al Sallab and M. A. Rashwan, "E-mail classification using deep networks," *Journal of Theoretical and Applied Information Technology,* vol. 37, pp. 241-251,2012.

[64] J. Pujara, H. Daumé Iii, and L. Getoor, "Using classifier cascades for scalable e-mail classification," in *8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, CEAS 2011*, Perth, WA, 2011, pp.55-63.

[65] N. Chatterjee, S. Kaushik, S. Rastogi, and V. Dua, "Automatic email classification using user preference ontology," in *International Conference on Knowledge Engineering and Ontology Development, KEOD 2010*, Valencia, 2010, pp.165-170.

[66] D. M. Jones, *Learning to Improve E-mail Classification with numero interactive*. Godalming: Springer-Verlag London Ltd, 2010.