SURVEY OF E-MAIL CLASSIFICATION: REVIEW AND OPEN ISSUES

Ekhlas Ghaleb Abdulkadhim[1]
[1] *Collage of Tourism Sciences*
University of Kerbala
Kerbala,Iraq
ekhlassghaleb@gmail.com

Muqdad Abdulraheem Hayder [2]
[2]*College of Education for Human Sciences*
University of Kerbala
Kerbala,Iraq
mukdadabdulraheem@gmail.com

Abstract **- Email is an economical facet of communication, the importance of which is increasing in spite of access to other approaches, such as electronic messaging, social networks, and phone applications. The business arena depends largely on the use of email, which urges the proper management of emails due to disruptive factors such as spams, phishing emails, and multi-folder categorization. The present study aimed to review the studies regarding emails, which were published during 2016-2020, based on the problem description analysis in terms of datasets, applications areas, classification techniques, and feature sets. In addition, other areas involving email classifications were identified and comprehensively reviewed. The results indicated four email application areas, while the open issues and research directions of email classifications were implicated for further investigation.**

Index Terms - Machine Learning Techniques, Email Classification, Spam Detection, Multi-folder Categorization, Phishing Detection.

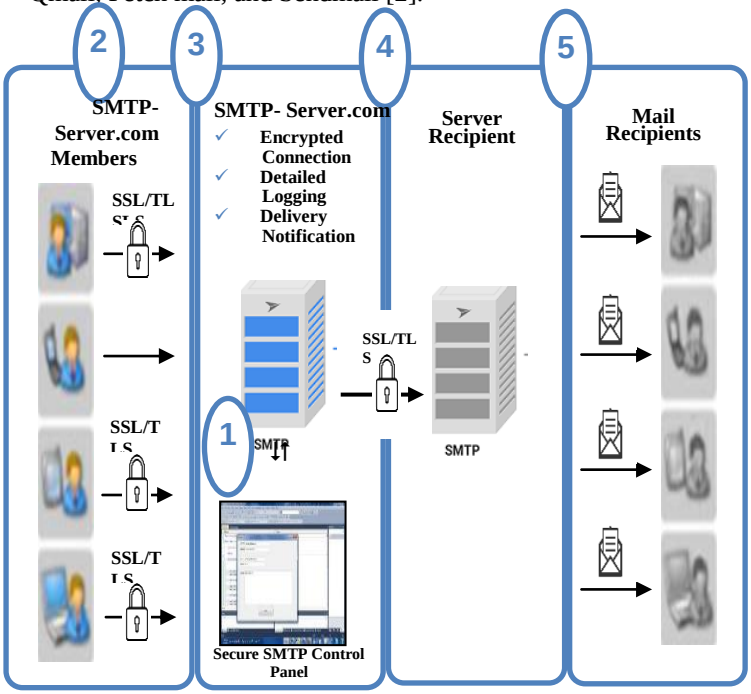
I. INTRODUCTION

Email is an economical and potent facet of communication, which has remarkably affected the personal and professional life of the modern human. However, there have been various cases of email misuse in the form of computer malware and spams, which are perpetrated via email and sent to the users' inbox as unwanted information. According to the reported statistics in 2014, 54 billion spam emails are sent to users per day on average.

Spam emails are predominantly mercantile or have attractive links to popular websites, while they connect the user to meddlesome domains, which diminish privacy, spread viruses, occupy space in the email box, and destroy the email servers. Consequently, substantial time is wasted in the filtration of import email and cancellation of unwanted emails. The classification of the email problems in this regard has led to the terms 'spam' or 'non-spam' to show the propriety of email messages [1].

A. How Does Email Work?

The simple mail transfer protocol (SMTP) is used to send emails in the form of plaintext via network throughout the world . In addition, error reporting or messaging and extra authentication could be attached to emails as user demand for advanced email grows. In this process, mail transfer agents (MTAs) are run secondarily to transfer messages between hosts to allow the mailing of messages across different countries. MTAs could arrive via software such as Postfix, Qmail, Fetch mail, and Sendmail [2].



Fig, 1. Email working technique

In this regard, the SMTP protocol per computer allows the passing of the message to the end address accurately. Switching on the light causes numerous emails to be sent to the hosts regardless of their regions without difficulty. Figure 1 depicts the simple function of an email sent from the sender to the recipient by the SMTP.

1. alice@yahoo.com sends out an email to bob@gmail.com;
2. Alice's email is received by the MTA at www.yahoo.com and queued (waiting lists) for delivery after the other messages that are also ready to be sent.
3. On port 24, MTA www.yahoo.com meets MTA www.gmail.com. After the connection is confirmed by www.yahoo.com, the message is sent out by MTA at www.gmail.com and accepted by www.yahoo.com, which confirms the reception of the message and discontinues the connection.
4. The message is placed inside Bob's incoming mailbox by the MTA www.gmail.com, and when Bob logs in, the presence of a new message is announced.

Evidently, the mentioned process could face some complications. For instance, if the user is not at www.gmail.com, the MTA at www.gmail.com rejects the message, reporting the issue to the MTA at www.yahoo.com, and the MTA at www.yahoo.com produces and sends a message to alice@yahoo.com, informing the absence of Bon (sender) at www.gmail.com.

In another hypothetical situation, www.gmail.com may not respond to the connection attempts of www.yahoo.com due to the fact that the host is off for maintenance or repair for instance. Under such circumstances, the MTA at www.yahoo.com informs Alice that the first delivery attempt has been problematic. As a result, the server manager determines more attempts at specific intervals until the deadline is met, and Alice will be informed that the message cannot be delivered. Recently, security measures and protocols have been developed for safe email transfer [2,3].

II. Problem Statement

Automatic email classification is considered to be the foremost means to the management of emails. In this approach, an email classifier system is applied for the automatic classification of emails into several specific sets of predefined categories. Figure 2 illustrates the structure of an automatic email classifier system. As can be seen, email classification occurs on three levels of classification, learning, and pre-processing. Initially, the automatic email classifier is propagated by the collection of an email dataset. For instance, an automatic spam email classifier could be developed by collecting a spam email dataset, which should contain both spams and non-spams for the training of the classifier. Following the collection of the dataset, the dataset should be cleaned; this is structurally known as data pre-processing in email classification that is automatic. In this stage, unnecessary or stop words are also eliminated to diminish the data volume and examine their dispositions. Furthermore, the pre-processing stage involves the stemming and lemmatization of token words and their conversion into the original form (e.g., 'exhibiting' to 'exhibit').

In this context, learning is a stage that encompasses the development of feature sets and feature extraction. In this context, 'feature' refers to the signs representing specific aspects of the activity or behaviours of the users that should be assessed. When it comes to email classification, proper feature set extraction is considered critical to increase the efficacy and accuracy of the learning tasks. When the features are extracted, the most distinguished features are considered for the classification process and improvement of the function of the classifier in terms of efficacy and accuracy. Notably, the construction and saving of the classifier is aimed at the classification of incoming emails. At the final stage of classification, the incoming emails are classified by the constructed classifier into specific categories (e.g., ham, spam, phishing) [1,4].

Several authentication mechanisms are employed by mail server engines for the analysis of email content and email classification as ham/spam or phishing/legitimate in the form of white or black lists, and these approaches could be optimized by the users. White and black lists are utilized for the comparison of the sources of new emails with the database in order to determine whether they should be classified as spam. On the other hand, emails are filtered by an alternative method, which involves feature extraction from the emails by classification approaches such as the Naïve Bayes algorithm, random forest algorithm, support vector machine (SVM), and neural networks.

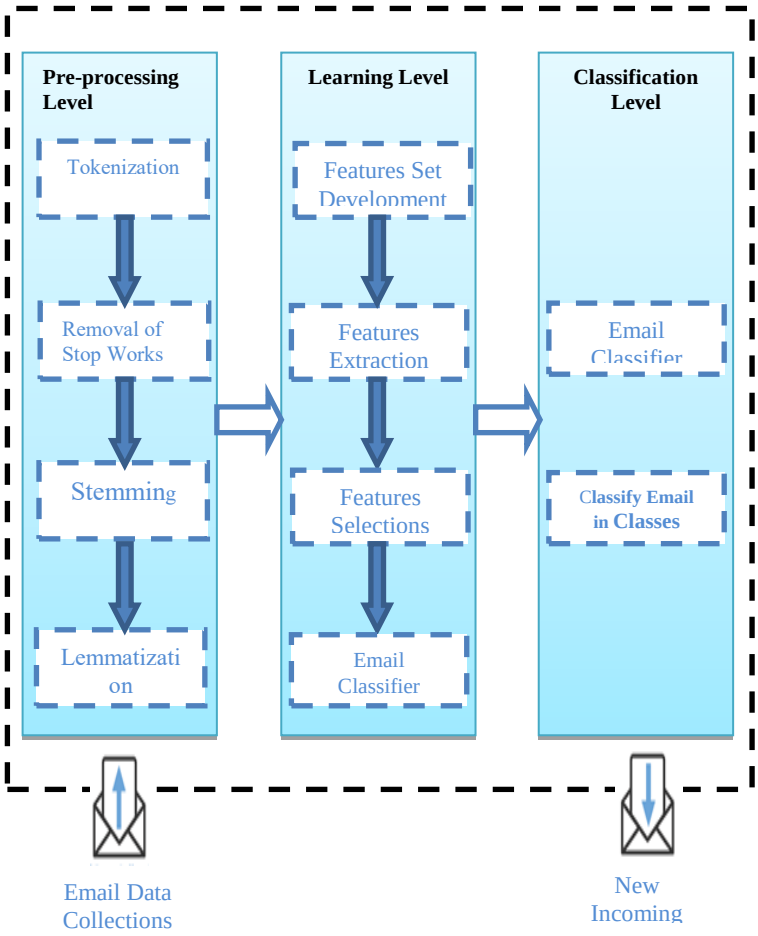


Fig. 2. General architecture of automatic email classification

In the majority of the studies in this regard, email classification has been performed based on the terms occurrence of the email. On the other hand, few studies have evaluated the semantic features of textual emails. According to the reported findings, the integration of semantic features with email classification approaches could expand the benefits of improves computational function and classification accuracy. Presently, experts are concerned with email classification to categorize spam (ham) emails into legitimate (phishing) emails. A small number of literature reviews have been

focused on spam and phishing email classification in terms of text classification. Such an example is a survey conducted in 2017 reviewed 98 articles published during 2006-2016, indicating that email classification has been employed in 15 arenas [1]. In simple terms, these arenas were classified into the dimensions of phishing, spam, multi-folder categorization, spam/phishing, and others. The present study may be considered an extension of the mentioned review study, providing researchers with the opportunity to comprehensively assess the applied methodologies and obtained results in this regard. Our review was performed on 40 articles published during January 2016-2020, which were retrieved from the core collection available on Web of Science and Scopus. Furthermore, this review study could aid spam email classification scholars based on the following issues that will be addressed in the following sections:

1. The determination of the application arenas of email classification;
2. The determination of the publicly accessible datasets for use in email classification;
3. The determination of the frequent features for use in email classification;
4. The determination of the frequently applied machine learning techniques in email classification;
5. The determination of the performance evaluation metrics for the assessment of email classifier function;

Application Arenas in Email Classification

According to the results of the mentioned review study, email classification was utilized in 15 arenas since 2006 until the beginning of 2016. The foremost arenas were shown to be phishing, spam, multi-folder categorization, spam/phishing, email thread, Chinese spam email detection, complaint email classification, and inquiry. Since five years ago, research has mostly been focused on the four arenas of phishing, spam, multi-folder categorization, and spam/phishing as denoted in our research (Figure 3).

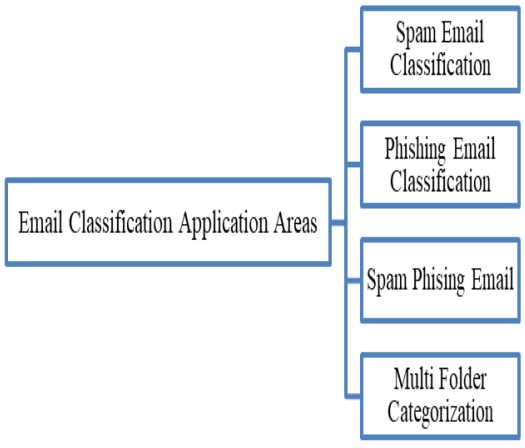


Fig. 3. Application areas in email classification

III. Email Classification Dataset

This section contains the analyzed datasets used in email classification. Table I shows the detailed analysis of the applied datasets in these arenas, as well as the dataset name, number of the studies in this regard, and the references regarding the use of a specific dataset. According to the obtained results, the PU dataset has most frequently been used in spam email classification as these emails are retrieved from the emails that have been exchanged between senders and receivers.

According to the findings, Phishing Corpus is the most commonly applied dataset used for phishing email classification, which consists of a set of hand-screened emails. In addition, Phishing Corpus has been employed in phishing and spam email classification (phishing emails), while a combination of PU, Ling Spam, Spam Assassin, TREC, and Spam Base datasets has been utilized for spam detection.

Multi-folder categorization has been performed using the Enron email dataset owing to the availability of the largest dataset for email classification. However, the Enron spam corpus differs from Enron email datasets as the former is a replacement for Ling Spam, Enron email dataset, and PU as elaborated in the previous studies in this regard. Notably, customized datasets have been frequently used by the researchers of email classification.

TABLE I
DETAILED ANALYSIS OF THE APPLIED DATASETS IN ALL AREAS OF E-MAIL CLASSIFICATION.

S.No.	Dataset Name	No.of Studies	Ref.
1	UCI	5	5,12,14,21,24
2	SpamAssasin	9	2,4,5,13,15,22,25,31,33
3	LingSpam	2	2,3
4	TREC	1	2
5	Nazario	2	22,33
6	Enron	8	7,8,10,17,24,31,33,38
7	SpamBase	6	1,3,11,14,16,17
8	PhishingCorpus	2	25,37
9	Phishing E_mail	1	26

IV. Feature Extraction and Feature Selection

The email activity or behaviour of users is described by feature. Feature extraction and selection play a pivotal role in the development of accurate and efficient classifiers in email classification systems [1].

Feature extraction could effectively enhance the email classification process. Extracted features include a set of objects and expirations that convert images into text, thereby determining whether the email is harmful. Prior to feature extraction, email pre-processing must be performed on all the emails through the reduction of high dimensionality (e.g., HTML tags, URL, email addresses). Pre-processing facilitates the feature extraction of emails. In the feature extraction process, spam and non-spam words are distinguished, with the

spam words verified through the selection of all the words that have been repeated at least 100 times in spam emails and added to the word list. At the next stage, an index is created of the words mentioned in the email and word list based on the list of the word indices. Finally, each email is converted into a vector as jth is the word in the word list, and the yj feature is equal to one if the jth word is found in the email and the yj feature is equal to zero, and if the jth word is not found in the email. This process is known as the vector space model based on binary weights .

Feature selection in the other hand is applied to develop a new structure from a set of essential features with the aim of reducing the dimensions of the search area and selecting high-weight features. Among the common methods of feature selection are the wrapper approach and filter approach, with the latter independent of the machine learning technique and more cost-efficient than the former. On the other hand, the feature subset could be estimated by the wrapper approach based on the machine learning technique, thereby providing better outcomes than the filter approach in the case of some issues. Additionally, the wrapper approach for feature selection has proven to yield better classification in the case of spam emails.

According to the results of the present study, the most common features in the context of the study include email URL, body, JavaScript, header, Spam Assassin, term-based, network-based, stylometric, online/offline, phrase-based, rule-based, concept-based, and social, structural or lexical features. Figure 4 depicts the taxonomy of these features based on the corresponding email classification application arenas. The overview of these features has been presented in the following section [1,4,5].

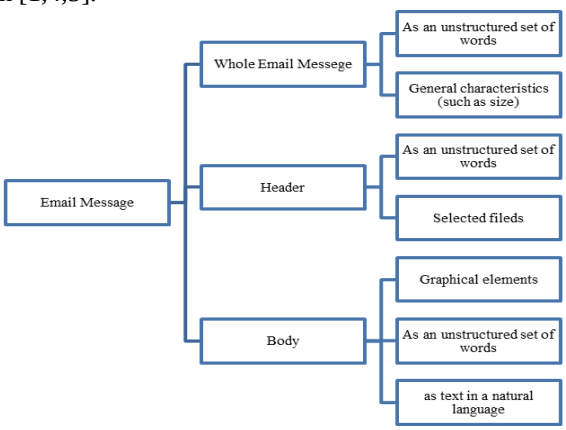


Fig. 4. Taxonomy of email features

V. Email Classification Techniques

There are five categories of email classification, including unsupervised, semi-supervised, and supervised machine learning, as well as statistical and content-based learning (Figure 5). In supervised machine learning, there are input instances incorporated into the learning algorithm although the output labels may fail to accurately identify a function roughly demonstrating this generalized behavior. Supervised learning techniques may be proposed in the form of the Naïve Bayes algorithm, SVM, artificial neural networks, and genetic algorithm.

In unsupervised machine learning, the learning algorithm has input instances, while the similar patterns in the input instances are identified by the output labels for the detection of an output (e.g., K-means algorithm-based clustering). On the other hand, the semi-supervised format refers to the supervised format with minor, labeled data without the need for major labeled data; active learning is an example in this regard. Keywords in emails are used for classification in content-based techniques. As for statistical learning, score or probability is assigned to each keyword , with the incoming emails are classified based on the total score or probability [6,7].

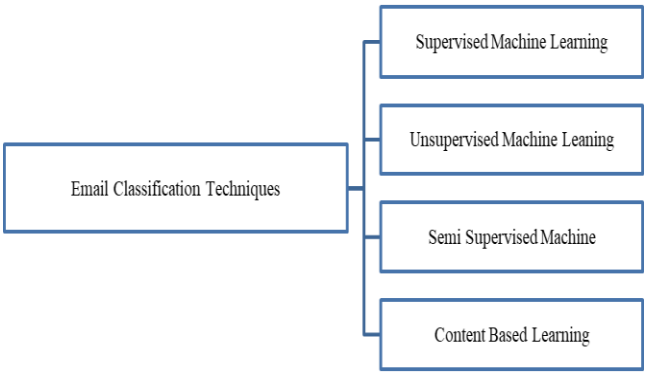


Fig. 5. Types of email classification techniques

VI. Literature Review

The aforementioned techniques have been applied in the previous studies in this regard, while supervised machine learning has been used most frequently. Table II depicts the distribution of articles based on various application arenas. Table III shows an overview of email classification techniques, with the data categorized based on the types of the email classification methods and each row containing the name of the techniques and number of the studies focused on multi-folder categorization, phishing, spam classification, and spam/phishing, as well as the references of each. According to the literature review, 21 out of the 40 studies in this regard have employed the spam classification technique, while 12 studies have employed the phishing method.

TABLE II
DISTRIBUTION OF ARTICLES ACCORDING TO APPLICATION AREAS

Se.No.	App.area	No.of Studies	Ref.
1	Spam	21	[2- 22]
2	Phishing	12	[23-34]
3	Spam & Phishing	3	[35-37]
4	Multifolder Categories	4	[38-41]

TABLE III
SUMMARY OF E-MAIL CLASSIFICATION TECHNIQUES

S.No.	App. Area	Approach	Author	Algorithm	Dataset used	Ref.
1	Spsm Classification	supervised	Aja.S	Bayesian and SVM	Spambase	[2]
2		content-based	Yılmaz.K	hifted-1D-LBP	Spamassasi, TREC	[3]
3		supervised	Sanjiba.S	ELM and SVM	Spambase	[4]
4		supervised	Hossam.F	Forest Classifier	SpamAssassin	[5]
5		Content-Based	Ali Rodan	(BBO)	SpamAssassin,UCI	[6]
6		supervised	Mis.Elifenes	BL,KNN,SVM	Unstructured	[7]
7		supervised	Anj.R	NBes, J48 DT	Enron	[8]
8		supervised	Sakha.A	SVM,Extra-Trees	Enron.Avocado	[9]
9		supervised	Ahmed.A	Artificial NN	Statistical info.	[10]
10		supervised	Eman M.	RF, RBF,SVM , J48	Enron	[11]
11		supervised	Shafi'I	BLR,HNB,RF	Spambase	[12]
12		supervised	Prachi.M	Naive Bayes	UCI	[13]
13		supervised	Hossam. F	(GA) and (RWN)	SpamAssassin	[14]
14		supervised	M.Bassiouni	RF, ANN, SVM,RT, KNN,	Spambase UCI	[15]
15		supervised	Amany A.	SVM,KNN	SpamAssassin	[16]
16		Content Based	Lamiaa M.	Pegasos algorithm	spambase	[17]
17		Content Based	el Bakrawy	WOA,rotation forest	Spambase,Enron	[18]
18		unsupervised	Maryam	NB,(ANN), SVM	unstructured	[19]
19		Content Based	Mi ZhiWei	J48,Chi Square	Huang &Chen,	[20]
20		Semisupervisd	Yeqin Shao	Active Clustering	unstructured	[21]
21		supervised	Dima.S	RF, Naïve Bays,	UCI Learning	[22]
22	Phishing Classification	supervised	Adwan.Y	RF ,J48	spam assassin	[23]
23		supervised	Junaid.A	cyber security	unstructured	[24]
24		supervised	Dr. Nang.S	(SVM),(FS)	Enron, UCI	[25]
25		supervised	Naghmeh.M	Neural Network	SpamAssassin	[26]
26		unsupervised	H. S. Hotaa	RRFST algorithm	phishing E-mail data	[27]
27		supervised	AnuVazhayil	Random Forest	TDM	[28]
28		Semisupervisd	Shelby R.	A custom algorithm	unstructured	[29]
29		supervised	Hiransha M	Keras Word,NN	unstructured	[30]
30		supervised	Harikrishnan	Rt, NB,DT, SVM.	unstructured	[31]
31		supervised	YONG.F	(RCNN) model	Enron,SpamAssass.	[32]
32		Semisupervisd	Hiba.Z	active learning	Custom dataset	[33]
33		supervised	LukášH	RNNs	SpamAssas +Enron	[34]
34	Spam & phishing	supervised	Prajakta.P	SVM	Custom dataset	[35]
35		supervised	Muhamet.B	Bayesian algorithm	15 set of dataset	[36]
36		supervised	Nidhin A	NB, DT, K NN, RF, SVM	(IWSPA-AP 2018)	[37]
37	Multifolder categorization	supervised	Rogerio.B	NB and SVM	corpus	[38]
38		unsupervised	Nesara.M	K-means Clustering	Enron	[39]
39		supervised	Aston.Z	Neural network	experimental data	[40]
40		unsupervised	Aaknksha.S	LDA	experimental data	[41]

According to the information in Table III, the studies regarding email classification techniques have primarily categorized emails as spam or ham. Among the 40 articles, 21 cases were focused on spam email classification, and most of the remaining studies (n=12) evaluated email classification phishing by developing binary classifiers to categorize emails as phishing or ham. The other related articles (n=4) assessed the multi-folder categorization

of emails by developing multi-class classifiers to categorize emails as user-defined email directories. Finally, three articles evaluated spam and phishing email classification by developing ternary classifiers for email classification as spam (ham) or phishing. Moreover, recent studies have categorized spam emails by image-based and text-based features. Table II shows distribution of the application arenas in detail with the related references.

The summary of the email classification techniques is presented in Table III and Figure 6 illustrates numbers of studies for e-mail classification techniques. In the present study, a total of 40 articles were selected and reviewed from the Web of Science core collection and Scopus database. In 28 studies, supervised learning was applied, while content-based techniques were applied in five studies, unsupervised machine learning was used in four studies, and semi-supervised machine learning was employed in three studies. According to our findings, SVM has been most commonly applied in supervised machine learning (12 studies), followed by the random forest algorithm (nine studies), neural networks (eight studies), Naïve Bayes algorithm (six studies), decision tree algorithm (four studies), and J48 (four studies). Semi-supervised machine learning was only observed in three articles, as SVM algorithm with active learning and the voting algorithm with active learning and were also applied in these studies. In addition, unsupervised techniques were applied in four studies, three of which also benefited from the K-means clustering technique. Among the 40 retrieved articles, content-based learning was reported in five cases.

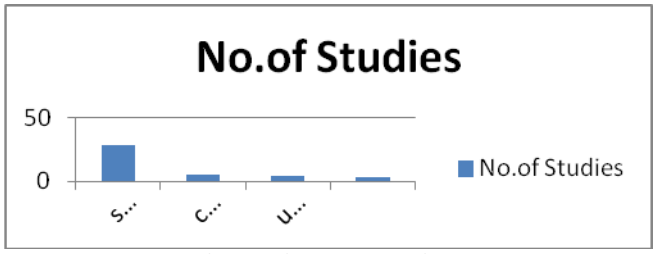


Fig. 6. No. of studies for e-mail classification techniques

CONCLUSION

The results of this study contribute to the research regarding email classification through the comprehensive analysis of the related studies published during 2016-2020. We exploited description analysis in the four dimensions of application arenas, datasets, features sets, and classification techniques in 40 articles, which were meticulously retrieved and evaluated. Analytically, the findings of the reviewed studies showed five main application arenas for the classification of email, including spam, multi-folder categorization, phishing, and spam/phishing. Furthermore, the main approaches to email classification were determined to be supervised, semi-supervised, and unsupervised machine learning, as well as content-based learning. According to the obtained results, supervised machine learning has been most frequently used, with SVM having the highest applicability and providing outcomes with higher accuracy compared to the other techniques in this regard. The comparison of SVM with the other approaches also demonstrated that it could yield better outcomes based on the features that are accessed through the master feature vector since it has no risk of over-fitting.

REFERENCES

[1] Mujtaba, Ghulam, et al. "Email classification research trends: Review and open issues." IEEE Access 5 (2017): 9044-9064.

[2] Sharma, Ajay, and Anil Suryawanshi. "A novel method for detecting spam email using KNN classification with spearman correlation as distance measure." International Journal of Computer Applications 136.6 (2016): 28-35.

[3] Kaya, Yılmaz, and Ömer Faruk Ertuğrul. "A novel approach for spam email detection based on shifted binary patterns." Security and Communication Networks 9.10 (2016): 1216-1225.

[4] Roy, Sanjiban Sekhar, and V. Madhu Viswanatham. "Classifying spam emails using artificial intelligent techniques." International Journal of Engineering Research in Africa. Vol. 22. Trans Tech Publications, 2016.

[5] Faris, Hossam, Ibrahim Aljarah, and Bashar Al-Shboul. "A hybrid approach based on particle swarm optimization and random forests for e-mail spam filtering." International Conference on Computational Collective Intelligence. Springer, Cham, 2016.

[6] Rodan, Ali, Hossam Faris, and Ja'far Alqatawna. "Optimizing feedforward neural networks using biogeography based optimization for e-mail spam identification." International Journal of Communications, Network and System Sciences 9.01 (2016): 19.

[7] Yitagesu, Mis Elifenesh, and Manisha Tijare. "Email classification using classification method." International Journal of Engineering Trends and Technology (IJETT) 32.3 (2016): 142.

[8] Radhakrishnan, Anju, and V. Vaidhehi. "Email Classification using Machine learning algorithms." International Journal of Engineering and Technology 9.2 (2017): 335-340.

[9] Alkhereyf, Sakhar, and Owen Rambow. "Work hard, play hard: Email classification on the Avocado and Enron corpora." Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing. 2017.

[10] Alghoul, Ahmed, et al. "Email Classification Using Artificial Neural Network." (2018).

[11] Bahgat, Eman M., et al. "Efficient email classification approach based on semantic methods." Ain Shams Engineering Journal 9.4 (2018): 3259-3269.

[12] Shuaib, Maryam, et al. "Comparative analysis of classification algorithms for email spam detection." International Journal of Computer Network and Information Security 10.1 (2018): 60.

[13] Mahajan, Prachi, Snehal Bhoite, and Abhijit Karve. "Intelligent Spam Detection Micro service With Server less Computing." (2018).

[14] Faris, Hossam, et al. "An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks." Information Fusion 48 (2019): 67-83.

[15] Bassiouni, Mahmoud, M. Ali, and E. A. El-Dahshan. "Ham and spam e-mails classification using machine learning techniques." Journal of Applied Security Research 13.3 (2018): 315-331.

[16] Naem, Amany A., Neveen I. Ghali, and Afaf A. Saleh. "Antlion optimization and boosting classifier for spam email detection." Future Computing and Informatics Journal 3.2 (2018): 436-442.

[17] el Bakrawy, Lamiaa. "Hybrid Particle Swarm Optimization and Pegasos Algorithm for Spam Email Detection." Advances in Systems Science and Applications 19.3 (2019): 11-22.

[18] Shuaib, Maryam, et al. "Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification." SN Applied Sciences 1.5 (2019): 390.

[19] Sharaff, Aakanksha, and Naresh Kumar Nagwani. "Identifying categorical terms based on latent Dirichlet allocation for email categorization." Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. 431-437.
[20] ZhiWei, Mi, Manmeet Mahinderjit Singh, and Zarul Fitri Zaaba. "Email spam detection: a method of metaclassifiers stacking." The 6th international conference on computing and informatics. 2017.
[21] Shao, Yeqin, et al. "A hybrid spam detection method based on unstructured datasets." Soft Computing 21.1 (2017): 233-243.
[22] Suleiman, Dima, and Ghazi Al-Naymat. "SMS spam detection using H2O framework." Procedia computer science 113 (2017): 154-161.
[23] Yasin, Adwan, and Abdelmunem Abuhasan. "An intelligent classification model for phishing email detection." arXiv preprint arXiv:1608.02196 (2016).
[24] Chaudhry, Junaid Ahsenali, Shafique Ahmad Chaudhry, and Robert G. Rittenhouse. "Phishing attacks and defenses." International Journal of Security and Its Applications 10.1 (2016): 247-256.
[25] Mon, Ei Ei, and Nang Saing Moon Kham. "Studying Email Filtering Approach to Identify Spear Phishing Attacks." International Conference on Computer Applications. 2016.
[26] Moradpoor, Naghmeh, Benjamin Clavie, and Bill Buchanan. "Employing machine learning techniques for detection and classification of phishing emails." 2017 Computing Conference. IEEE, 2017.
[27] Hota, H. S., A. K. Shrivas, and Rahul Hota. "An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique." Procedia computer science 132 (2018): 900-907.
[28] Vazhayil, Anu, et al. "PED-ML: Phishing email detection using classical machine learning techniques." Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA). Tempe, AZ, USA, 2018.
[29] Curtis, Shelby R., et al. "Phishing attempts among the dark triad: Patterns of attack and vulnerability." Computers in Human Behavior 87 (2018): 174-182.
[30] Hiransha, M., et al. "Deep learning based phishing e-mail detection." Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA). Tempe, AZ, USA, 2018.
[31] Harikrishnan, N. B., R. Vinayakumar, and K. P. Soman. "A machine learning approach towards phishing email detection." Proceedings of the Anti-Phishing Pilot at ACM International Workshop on Security and Privacy Analytics (IWSPA AP). Vol. 2013. 2018.
[32] Fang, Yong, et al. "Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism." IEEE Access 7 (2019): 56329-56340.
[33] Zuhair, Hiba, and Ali Selamat. "Phishing classification models: issues and perspectives." 2017 IEEE Conference on Open Systems (ICOS). IEEE, 2017.
[34] Halgas, Lukas, Ioannis Agrafiotis, and Jason RC Nurse. "Catching the Phish: Detecting Phishing Attacks using Recurrent Neural Networks (RNNs)." arXiv preprint arXiv:1908.03640 (2019).
[35] Patil, Prajakta, Rashmi Rane, and Madhuri Bhalekar. "Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm." 2017 International Conference on Inventive Systems and Control (ICISC). IEEE, 2017.
[36] Baykara, Muhammet, and Zahit Ziya Gürel. "Detection of phishing attacks." 2018 6th International Symposium on Digital Forensic and Security (ISDFS). IEEE, 2018.
[37] Unnithan, Nidhin A., et al. "Detecting phishing E-mail using machine learning techniques." 51-57.
[38] Bonatti, Rogerio, et al. "Effect of part-of-speech and lemmatization filtering in email classification for automatic reply." Workshops at the Thirtieth AAAI Conference on Artificial Intelligence. 2016.
[39] Madhav, Nesara, et al. "Email Categorization Advisor for Help Desk." Perspectives in Communication, Embedded-systems and Signal-processing-PiCES 1.5 (2017): 66-70.
[40] Zhang, Aston, et al. "Email category prediction." Proceedings of the 26th International Conference on World Wide Web Companion. 2017.
[41] Sharaff, Aakanksha, and Naresh Kumar Nagwani. "Identifying categorical terms based on latent Dirichlet allocation for email categorization." Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. 431-437.