IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

E-mail Classification System: A Review and Research Challenges

Aruna Kumara B[1*], Mallikarjun M Kodabagi[2]

School of C & IT, REVA UNIVERSITY, Bengaluru, India

Corresponding Author: arunakumara.b@reva.edu.in Tel.: 0-8951755795

DOI: https://doi.org/10.26438/ijcse/v7si14.489495 | Available online at: www.ijcseonline.org

Abstract — Individuals and corporate user's appetite to use email as one of the vital sources of communication. Email has become one of the part and parcel of our lives. Due to globalization, there is an extensive increase in the volume of emails received by a user. A particular user receives about 50-60 emails per day of different categories, for some users it may reach 100 emails. Out of these emails, most of them are not related to user interest. As the volume of emails receive continues to grow, the user has to spend a significant amount of time to process emails. It requires a system to manage these emails and to develop an automated classification system to classify emails into various categories as per the individuals and professional needs such as: academic, business, commercial. This paper presents a comprehensive review of several articles of email classification. The generic framework for email classification is devised and various steps in the framework are discussed in detail. The comparative analysis of various email classification techniques is discussed. The various challenges in the field of email classification are also presented.

Keywords— E-mail classification, E-mail categorization, Text classification, Preprocessing techniques, Feature extraction and Machine learning techniques.

I. INTRODUCTION

Information or data is the heart of business processes and the decision makers could make use of data to add appreciated intuition to business. The volume of data storing in the electronic format is growing exponentially every year. Email is one of the most widely used platforms to exchange information in the electronic format for communication purpose. In recent years, usage of email increasingly occupies a significant position in the field of information exchange such as academics, corporate, business and commercial purpose. A particular user receives around 50–60 emails per day [1]; for other users who uses email regularly, hundreds of emails are habitual and most of them are not useful. Users have to be compelled to pay a vital a part of the operational time on handling emails. Consequently, management of such emails could be a concern encountered by administrations and individual users, and it imposes the requirement to formulate ways that logically wear down the matter. Usually, the focal tool to manage such types of emails is classifying the emails automatically based on their category [2] & [3]. A framework of an automatic email classifier, classifies emails into a distinct set of predefined classes automatically. For instance, for email management, an automatic email classifier framework that classifies an incoming email into personal, or official (i.e., based on the requirements of an individual), and many more.

Some of the applying areas wherever email classification are often applied are as follows; multi-folder categorization (i.e., classifying an incoming email into different classes such as exam, circular, research, placements and academics in the academics field), spam email classification, text and image-based email classification, phishing email classification, etc.

This review might give help to the researchers operating within the field of e-mail classification. The details of various works discussed such as: pre-processing techniques used to perform data pre-processing, widely used feature sets to identify the class of an incoming email, different machine learning techniques used to classify an email into different classes and various performance metrics used to assess the performance of email classifier system are collated and presented.

The paper is ordered as follows; Section 2 reviews use of various pre-processing techniques, use of assorted feature sets and the analysis of various machine learning techniques employed in email classification. Section 3 gives the generic framework for e-mail classification method. Section 4 presents a comparative analysis of the work carried out. Section 5 presents some of the observations and research challenges. At the last, the conclusion of the work is presented in Section 6.

II. RELATED WORK

Some of the many contributions within the field of e-mail organization ar summarized within the following.

Pre-processing techniques plays a vital role while classifying e-mails into different categories. Uysal and Gunal [4] explores the impact of varied pre-processing tasks like converting data into lowercase, tokenization, removal of stop-words and stemming on text classification in terms of the many aspects like classification accuracy, text domain, text language, and dimension reduction. All probable combinations (i.e. 16 different combinations) of the pre-processing methods were considered for experimentation and assessed on 2 different domains, viz., news and email, and in 2 different languages viz. English and Turkish. The email dataset contains three hundred coaching and a hundred testing samples for every category, viz. spam and legit. Experiment is done on various feature sizes (10, 20, 50, 100, 500, 1000 and 2000) by considering all possible combinations of the previously mentioned pre-processing tasks and chi square based features were selected. The SVM classification algorithm is used and Micro-F1 score is used to measure the success.

The method reports that there was an affirmative impact on accuracy when lowercase conversion, tokenization and stemming applied and stop word removal was not applied. With respect to the domain and language analysis, it's not good to remove the stop word. For feature size analysis, the pre-processing tasks were selected which provides maximum scores at a minimum feature size. In e-mail domain, lowercase conversion was applied on both languages, and status of remaining pre-processing tasks varies liable to the language. In news domain, lowercase conversion and tokenization were applied In spite of the fact that there are numerous pre-processing techniques utilized in classification, the main lowercase transformation that improves exactness and there's no general combination provides winning classification results for every domain and language. Thus for a text classification drawback, a scientist ought to rigorously analyze all doable completely different combos of pre-processing tasks instead of fully disabling or sanctioning them.

Text categorization has rose to the level of need in identifying the document to which category it belongs to. In our daily life there are several substantial applications like distinguishing the genre of the text document becomes a lot of crucial in net applications. The work represented in [5] targeted on three issues in text classification of Turkish texts. The primary one is, to spot the author of an editorial. Second is, to see the gender of the author to classify text in line with the data and therefore the final one is to spot a class of a given text like sports, media, and economic science. This method uses n-gram model to classify text in terms of author, genre, and gender on documents of articles got from Turkish newspapers. A four steps method was used to extract n-gram features: stemming, punctuation removal, n-gram model and feature filtering. Also, authors used three different machine learning techniques viz., SVM, Naïve Bayes and Random Forest for text classification to predict the author, genre of the document and gender of the author. It is observed that three categories are selected for experimental purpose viz., sports, economics, and media. Five male and five female authors are selected for each category and collected their 10 recent articles. Also, the corpus consists of three hundred articles from totally different Turkish newspapers are collected, these are written by thirty totally different authors. F1 measure is employed to judge the performance of classifiers. This system divides the dataset into ten subsets, in every iteration, one set becomes the take a look at set and remaining becomes the coaching set. The trained model is evaluated on the take a look at set and its performance score is recorded. Finally, the common of all scores is taken as a final score of the model. The results report that genre and gender-based classification is done a lot of meritoriously than author primarily based classification. Character level N-Gram performs higher than word level. With reference to machine learning techniques, it's ascertained that SVM outperforms the Naïve Thomas Bayes and Random Forest. However, this methodology may use a most range of options within the word level the maximum amount as employed in character level, so the performance of the word level may increase.

There has substantial success seen in generating a skeleton [6] of frequent content based on earlier seen specimens for plain or structured data such as web pages. These structures can be convenient for tasks such as plagiarism detection, automatic labelling, duplicate detection and structured information extraction. An algorithm for template induction is developed which focus on plain text content. Construction of template consists of 2 parts: 1) clustering similar messages 2) for each cluster determining the parts which are considered "fixed" and store the information in a standard representation which is produces a template.

It is observed that using a suffix array is efficient than the using of shingling baseline for template induction. Also, the investigation shows that templatizing plain text using a suffix array is more efficient. This investigation suggests that, the generated templates could save 35 words on average while composing e-mails. However, this work did not address on a real-world scale of documents while collecting e-mail sizes. Also, it did not address the parallelization on both the clustering and induction, if so the performance could be improved.

Let us see how neural networks helps to categorize emails arrive during a fixed sized time window in the near future [7]. Machine generated messages or e-mails such as promotional campaigns, shopping receipts, booking confirmations, newsletters etc., are created by using a fixed template with little personalized information. Two types of neural networks are used in this work: 1) MultiLayer Perceptions (MLP) - a type of feed-forward neural network, 2) Long Short-Term Memory (LSTM) - a type of recurrent neural network. They considered machine generated emails as these type of emails contributes major percentage when considered all types of emails, and categorization can be done more accurately on them. Series of emails received by a user is considered and each email has a timestamp indicating when it was received. Authors developed an email categorization method and deployed at Google which is used to make categories like flight reservations, restaurant bookings, and event reminders. For each user in the training set, they considered the last 90 days of email messages received. For each email, record the category and the time it was received also it records the derived information of the time of receipt. To evaluate the different prediction approaches, the dataset is partitioned into two parts: the first 45 days of each user's history considered as a training set, the next 45 days history is considered as test and validation sets. K - dependent Markov chains method is used as a baseline system which is based on counting to calculate the probability of occurrence of an event after the observation of a chain of k-consecutive events. MLP network consists of neurons arranged in the form of layers and each one is connected to another. The top and bottom layers form the input and output layer. K-received emails are provided as input, representing the category and time of receipt of each email. All three techniques are compared experimentally using historical emails of about 100,000 users and explored the effects of providing less or more history. It is observed that both types of neural network considerably outperforms k-dependent Markov chains. Under the best configuration, the success rate achieved is 0.8737 of 1, i.e., 87.37% of predicted top category emails will indeed arrive within 3 days. However, this study does not concentrate on more features of past emails, if so the prediction accuracy could be improved.

The huge data in the medical field makes researchers a rigid task to retrieve required information. Pre-processing techniques showed significant effect in text classification on MEDLINE documents [9] as well. This study assessed the result of mixing completely different pre-processing techniques with many classification algorithms conferred within the WEKA tool. The experiments showed that the appliance of pruning, stemming and word internet reduces considerably the amount of attributes and improves the accuracy.

The studies discussed above were worked on English language text only. The pre-processing techniques can be applied on other languages as well. The work [10] analyzes 3 reduction approaches that were connected on Arabic content. The methods embrace stemming, lightweight stemming and word bunches. The effect of the previously mentioned systems was examined and broke down on the K-nearest neighbor classifier. The examination metric incorporates the elements of report vectors, characterization time and exactness. Numerous tests appropriated exploitation four different representations of the corpus. The corpus comprises of 15000 archives that speak to 3 classes: sports, financial matters, and governmental issues. As far as vector sizes and grouping time, the stemmed vector devoured the most modest size and least time important to order a testing dataset that comprises of 6000 archives. The daylight stemmed vector outflanks the contrary 3 portrayals.

The machine learning algorithms are used to discover a pattern of monotonous keywords to classify emails as spam [11]. A model has been developed to classify the emails supported the parameters contained within the e-mail header like - To field, From field, cc/bcc field, Message-ID, etc., e-mail body with unremarkably used keywords and punctuations. Every parameter is taken into account as a feature once applied to a machine learning algorithmic rule model. This model is a pre-trained model to differentiate between an ambiguous and accurate output using feedback mechanism. The given model trains the algorithmic program and classifies emails from the antecedently classified dataset, later extends its practicality to classify incoming emails. From the experimental results, it is observed that productivity is increased and also distractions and clutters caused by spam were reduced.

A method to filter e-mails supported feature analysis combined with text classification is described in [12]. A Bayesian filtering methodology has used during this approach, it effusively employs and advances the basal technology of ancient spam filtering, defines the e-mail filtering define, and experiments in English information set. Experimental results showed that the given methodology is affordable and operative. However, this methodology doesn't work well for large-scale processing, testing for Chinese e-mails, learning a lot of inapplicable e-mail options, classifying supported 3 elements of the topic, body and, text-processed attachment, adding the understanding of the linguistics to e-mail text, and so on.

The study [13] explores the effect of mistreatment social network knowledge extracted from Associate in Nursing E-mail corpus to enhance detection of spam emails. The comparison of a social knowledge model with ancient spam knowledge models is performed by creating and assessing classifiers from each model varieties. From the results it is

observed that correct spam detectors may be produced from the low-dimensional social knowledge model alone, however, spam detectors generated from combos of the standard and social models were additional correct than the detectors produced from either model in isolation. However, this study might investigate OSN knowledge, which considers knowledge set consisting of a bigger range of messages connected to many OSN could yield additional reliable results.

Image spam is a type of spam in the field of spam detection where the details of advertisement are specified in the image. A method has been developed for classification of text and image based spam using artificial neural network [14]. Three machine learning algorithms used are KNN, Naive Bayes and reverse DB SCAN. Performance comparison of all three algorithms is provided based on four measuring factors namely: precision, sensitivity, specificity, and accuracy. The results showed that 50% of better accuracy is achieved for pre-processed data. KNN with pre-processing data achieved 83% accuracy compared to 45% accuracy without pre-processed data in both text and image-based spam. DBSCAN achieved 74% accuracy for pre-processed data compared to 48% without pre-processed data. Naïve Bayes achieved 87% accuracy for pre-processed data compared to 47% for unprocessed data. From the results, it is observed that Naïve Bayes achieved amazing accuracy for pre-processed data. However, text filtering method is time-consuming and text recognition was not always perfect. Also, this method was unable to predict CAPTCHA images.

This study [15] explored the performance of supervised machine learning techniques in real environments while most studies revealed the performance of datasets. The 3 different real environments considered are i) institute where research can be done ii) academics field in an university iii) company working for commercial purpose and over 1000 users. The results showed that classifiers performed poorer in the academic environment as the emails were more multifaceted and dissimilar. SVM and decision tree achieved better performance than the other classifiers. As the spam emails in the academic environment have more vibrant domains and contents, which makes SML classifiers more challenging to build an accurate model.

A fusion model of spam email classifier model is presented in [16]. To increase spam classification accuracy hybrid solution is used as the key algorithm approved by information gain calculation. They considered a model with three stages: pre-processing of email, extraction of features and classifying emails into various classes. The study reveals that implementation of spam filtering method on context – based emails is feasible. Key steps of the context-based email management begin with pre-processing email by POS Tagger later it extracts many features to transform emails into graphs later graphs are coordinated to the representative graph so emails are classified to the category with the best match represents. Linger implements information gain classifier for filtering spam and used a neural network to classify emails into uniform clusters. The results showed that 100 percent accuracy. However, this technique must think about reducing pre-processing time because the time taken for pre-processing in employing a spam filter and in not victimisation spam filter varies insignificantly.

III. GENERIC FRAMEWORK – E-MAIL CLASSIFICATION SYSTEM

The Generic framework for email classification mainly consists of 3 phases, namely: data pre-processing,

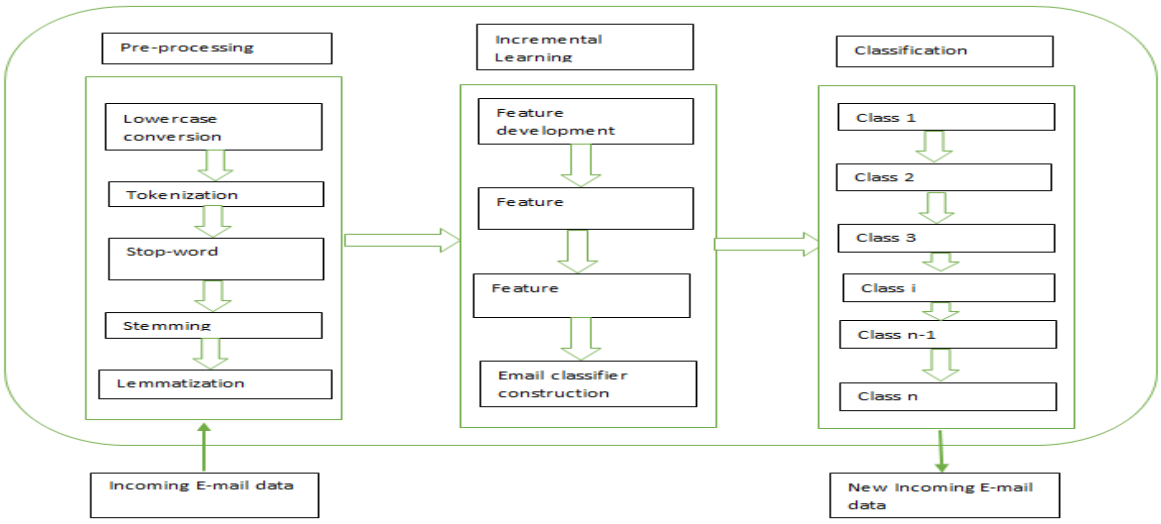

Fig. 1: Generic framework of E-mail Classification System

incremental learning and classification. Fig. 1 gives the generic framework of an automatic email classification system. To devise a framework to classify emails automatically, one has to collect data set first. For instance, if the objective is to build a framework for automatic email classification in the academic field, then one must collect private dataset who works in the academic field. The next task after data collection is data cleansing (i.e., data pre-processing). Data pre-processing [24] is one of the major phases in the knowledge discovery process. Raw data generally comes with many imperfections such as stop words, missing values, inconsistencies noise and/or redundancies. Performance of successive dynamic learning algorithms will be underperformed if they have submitted with low eminence data. Thus, by conducting proper pre-processing steps the quality of data can be enhanced and it can influence the reliability of successive automatic decisions [25]. In the data pre-processing phase, first, all the email data will be converted into a lowercase form, then the converted data will be converted into a token of words. This phase also expels unimportant information or stop words to lessen the size of information to be inspected for further procedure. At long last, stemming and lemmatization are connected on token of words to urge their root frames (e.g., "extracting" to "extract").

In the incremental learning phase, first the set of features are developed later they will be extracted. The term feature in email represents the behaviour or activity of an email for a specific user. In the email classification system developing a feature set and extracting useful features plays an essential and major role in making the incremental learning task more efficient. Once the features are extracted the most relevant features as per classification requirements are selected successively classification phase to improve the performance of the email classifier system.

Finally, an automatic email classifier system is constructed and saved in the classification phase. A constructed classifier framework is used to classify incoming emails into a defined category like academics, exams, placements, circulars, etc.

IV. COMPARITIVE ANALYSIS

This section gives analysis and assessment of various pre-processing techniques, feature sets and machine learning approaches used in different email classification methods.

Table 1 demonstrates the comparative analysis of different email classification systems and also represented graphically in Figure 2. X-Axis and Y-Axis in Figure 2

Table 1. Analysis of different E-mail Classification Techniques

Ref. number	Pre-processing techniques	Feature sets	Machine learning techniques	Accuracy
[4]	Lowercase conversion, Tokenization, Stop-word removal, Stemming (all possible combinations are used)	Feature sets based on chi square method	SVM	Lowercase conversion improves accuracy irrespective of domain and language.
[5]	Stemming, Punctual removal	Author, genre, gender	SVM, Naïve Bayes, Random Forest	SVM achieves 90% accuracy in genre based classification.
[7]	Stemming, Stop-word removal.	Features of past emails	Markov Chain	Under best configuration 87.37% accuracy is achieved.
[10]	Removal of punctuation, Removal of tags, Removal of stop-words, Stemming.	Sports, Economics, Politics	KNN	Stemmed vector achieves more accuracy.
[16]	Stemming	Informative features	Naïve Bayes, Hidden Markov Model (HMM)	HMM achieves 91.28%.
[21]	Stop-word removal	Sign-off words, greeting words and key words	Naïve Bayes, SVM	100% accuracy is achieved.

[22]	Tokenization	Semantic features	Semantic VSM (sVSM)	Accuracy achieved is 91%.
[23]	Tokenization, Stop-word removal	Size of the message, Length of the subject line, Attachments numbers, Type of attachments, Size of attachments, and Number of words present in the message.	Dis-agreement based semi-supervised learning	85.7% Accuracy is achieved.

represents the reference number and accuracy columns of Table1.

The comparative study reveals that, Lowercase conversion pre-processing technique improves classification accuracy when it was applied with the SVM machine learning approach. SVM approach achieves 90% accuracy with stemming and punctual removal preprocessing techniques. Markov chain machine learning technique achieves 87.37% with stemming and punctual removal pre-processing techniques [7]. HMM achieves 91.28 % accuracy, in which 100 important emails were considered for experimentation, in which 91 were classified correctly, 7 were classified incorrectly and 2 instances could not be resolute which gives the result [16].

The hybrid solution (SVM, Naïve Bayes) achieves 100% accuracy in spam email classification system with stop-word removal pre-processing technique [17]. The Semantic VSM (sVSM) achieves 91% classification accuracy, which is more than by 10% compared to other traditional methods [18]. A semi supervised machine learning technique based on disagreement achieves 85.7 % accuracy which is higher than the others by 3.4% [19]. This comparative analysis showed that stemming is the best among all the available data pre-processing techniques and SVM was the most frequently used machine learning technique. Also, this study reveals that 100% accuracy is achieved in spam email classification when a hybrid solution has been applied.

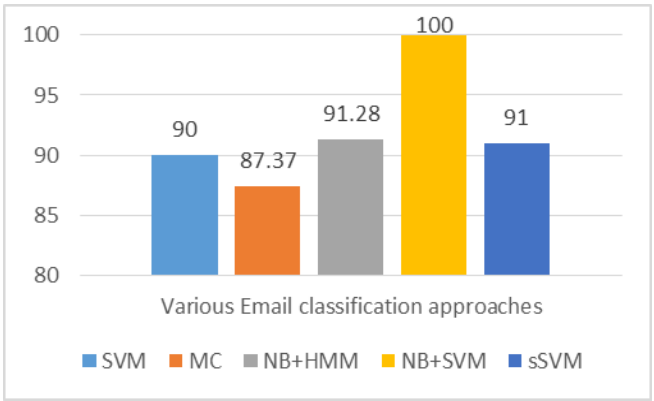

Fig. 2: Accuracy comparison between various Classification Techniques on different datasets

V. RESEARCH CHALLENGES

This segment highlights some research challenges in the field of email classification system:

(1) There is a requirement for auxiliary development of preprocessing techniques which are categorized by computational requirements for real-time decision making when big data and high-speed data streams are considered [21].

(2) Concept drift problem: Further research on feature selection methods is required which can directly address the concept drift problem [21].

(3) Real time learning: As per the observations made most of the present analysis on e-mail classification is done on the datasets that don't embody factors on real-time environment. Real-time environmental factors have an effect considerably on the performance classifiers. So, it's indeed to develop a real-time learning email classifier [1].

(4) Updating feature space dynamically: There is a need to develop the feature space which updates features dynamically without rebuilding the whole model [1].

(5) Deep learning for better classification: It permits computational approaches with numerous processing layers to learn representations of data with few dimensions of abstraction [5], [22].

(6) Hierarchical classification for Email classification: There is a need to develop a method which classifies email in a hierarchical way. There are complex classification issues in email classification system. To facilitate this issue, complicated problems within the email classification is deciphered by dividing them into many smaller tasks within which classifiers square measure developed in an exceedingly class-conscious manner. The first hierarchy with high level classification will be developed, for example, to find the received email belongs to spam category or other category, and low level classifiers will be developed by training them with different sub classes of high level classification. Finally, the low level classifiers are developed for a specific purpose.

VI. CONCLUSION

This investigation shows a thorough examination of different pre-processing techniques used and their impact on the email classification system. This survey also presents widely used feature sets and different machine learning techniques for the email classification system. The various performance metrics used to evaluate the email classification system are also described. The comparative analysis also presents various pre-processing techniques, datasets, feature sets, machine learning techniques, and performance metrics. This analysis showed that the most commonly used feature sets were header part, body part and behavioural part of the email. This collective analysis showed that stemming is the best among all the available data pre-processing techniques. Also, this survey showed that SVM was the most frequently used machine learning technique.

REFERENCES

[1] G Mujtaba et al.: Email Classification Research Trends: Review and Open Issues. IEEE Access, Vol. 5, 2017, pp. 9044-9064.

[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] Alper Kursat Uysal, Serkan Gunal: The impact of preprocessing on text classification. International Journal Information Processing and Management (50) 2014, pp. 104-112.

[5] Ayca Deniz, Hakan Ezgi Kiziloz. : Effects of various preprocessing techniques to Turkish text categorization using N-Gram features. IEEE 2nd International conference on Computer Science and Engineering (UBMK '17), 2017, pp. 655-660.

[6] Julia Proskurnia et al.: Template induction over Unstructured Email corpora. In: Proc. International World Wide Web Conference committee (IW3C2), 2017.

[7] 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.

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

[9] Carlos Adriano Gonc¸alves, Celia Talma Gonc¸ alves, Rui Camacho1, Eugenio Oliveira. : The impact of Pre-Processing on the Classification of MEDLINE Documents. 10th International workshop on pattern recognition in information systems, 2010, pp. 53-61.

[10] Rehab Duwairi, Mohammad Nayef Al-Refai, Natheer Khasawneh. : Feature reduction techniques for Arabic Text classification. Journal of the American Society for Information Science and Technology, 60(11):2347–2352, 2009.

[11] A. A. Alurkar et al.: A proposed data science approach for email spam classification using machine learning techniques. 2017 Internet of Things Business Models, Users, and Networks, Copenhagen, 2017, pp. 1-5.

[12] X. Li, J. Luo and M. Yin. : E-Mail Filtering Based on Analysis of Structural Features and Text Classification. 2010 2nd International Workshop on Intelligent Systems and Applications, Wuhan, 2010, pp. 1-4.

[13] A. Borg and N. Lavesson. : E-mail Classification Using Social Network Information. 2012 Seventh International Conference on Availability, Reliability and Security, Prague, 2012, pp. 168-173.

[14] A. Harisinghaney, A. Dixit, S. Gupta and A. Arora. : Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. 2014 International Conference on Reliability Optimization and Information Technology (ICROIT), Faridabad, 2014, pp. 153-155.

[15] W. Li and W. Meng. : An empirical study on email classification using supervised machine learning in real environments. 2015 IEEE International Conference on Communications (ICC), London, 2015, pp. 7438-7443.

[16] M. K. Chae, A. Alsadoon, P. W. C. Prasad and A. Elchouemi. : Spam filtering email classification (SFECM) using gain and graph mining algorithm. 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, 2017, pp. 1-7.

[17] 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.

[18] 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.

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

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

[21] S. R. Gomes et al.,: A comparative approach to email classification using Naive Bayes classifier and hidden Markov model. 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), Dhaka, 2017, pp. 482-487.

[22] Zhao Lu and Jianguo Ding. : An efficient semantic VSM based email categorization method. 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, 2010, pp. V11-525-V11-530.

[23] W. Li, W. Meng, Z. Tan and Y. Xiang. : Towards Designing an Email Classification System Using Multi-view Based Semi-supervised Learning. 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications, Beijing, 2014, pp. 174-181.

[24] S. Garcia, J. Luengo, F. Herrera. Data Preprocessing in Data Mining, Springer, 2015.

