An Intelligent System for Detecting Image Spam based on Exploration of Text Information Embedded into Images

Mallikka Rajalingam, M. Balamurugan

Abstract—Spam mails act as a immense hazard to communication security as it leads to phishing or virus attacks that could harm the user accounts and the organizations by exposing confidential information.With these inferences, the present research attempted to accurately detect image spam e-mails which have always been a topic of great research in data security. Detection of image spam mails is divided into two separate components- character segmentation and character recognition. While the former segments individual characters, the latter overcomes the issue of blocked texts and not well surveyed. The final phase of the work is a complete image spam detection system with the two proposed works built to detect spam messages. Various techniques have been proposed to tackle this setback and the principle of this paper is to evaluate and review several algorithms, talk about benchmark statistics or data, show appraisal of performance followed by potential direction for future work.

Keywords- Feature extraction, character segmentation, character recognition, image spam detection, and multi-SVM.

I. INTRODUCTION

Email correspondence is one of the proficient and most famous correspondence frameworks that empower individuals to speak with one another. The all out number of overall email accounts is relied upon to increment from 3.3 billion of every 2012 to 4.3 billion records before the finish of 2016 [1]. This speaks to a normal yearly development pace of 6% throughout the following four years. In such manner with such a disturbing utilization of email correspondence, overseeing messages against fake exercises has become a significant errand. One such movement through messages is the incautious presenting of undesirable email on clients acknowledged as spam mail. A unsolicited mail is characterized as a spontaneous/unimportant/undesirable E-mail communication got by clients [2]. Unsolicited email sends normally business or gainful battles of questionable items, dating administrations, easy money scams and promoting. Spam message furthermore spread vindictive or infection codes and is proposed for deceitfulness in monetary exchange or spoofing.

Junk e-mail is measured to control misfortunes more than the web particularly when they will in general twist malevolent for trade associations.

Revised Manuscript Received on January 06, 2020.

Mallikka Rajalingam, Department of Computer Science and Engineering, Bharathidasan University, Trichy, India. Email- mallikka2002@gmail.

Dr. M. Balamurugan, Department of Computer Science and Engineering, Bharathidasan University, Trichy, India. Email- comand mmbalmurugan**@gmail.com**

A few misfortunes are for the most part accidental losses not centering a specific system or any association. Spam sends possess more system transfer speed during transmission. It likewise expends client time as far as looking. Factual reports appear, since from December 2014, unsolicited or irrelevant mail represented 66.4 percent of email dealing wide-reaching and orient comprises 54 percent of the all out rate [3]. A ongoing investigation [4] uncovers the way that the greater part of the clients get additional junk messages than actual or legitimate messages.

Junk mail sends as a rule contain business or beneficial crusades of dubious items, dating administrations or easy money scams. A spam mail may also be denoted as Unsolicited Bulk Mail (UBM), spam mail, Unsolicited Commercial Email (UCE), junk mail, Excessive Multi-Posting (EMP) and bulk email (Cyberoam, n.d.) As spam mails often hold irrelevant and unwanted messages which may cause a collateral damage to the network, it is considered as a major threat to network security. Spam refers to an electronic equivalent of junk mails that are sent with an intention of trading products, services or promoting email scams. Most of the spam simply conveys the ideological thoughts of the purveyors with the idea behind to grab the attention of web users to their sites for simply making money.

II. PROPOSED IMAGE SPAM DETECTION TECHNIQUES

Picture based spam or picture spam is a sort of email junk where printed unsolicited message is inserted into images that are then joined to unsolicited messages. Since the vast majority of the email customers will show the picture record legitimately to the client, the junk message is passed on when the e-mail is open (there is no compelling reason to additionally open the joined picture document). The objective of picture unsolicited is plainly to dodge the investigation of the email's printed substance performed by a large portion of the junk channels (e.g., Spam Assassin, Radical Spam, Bogofilter, SpamBayes). Additionally, spammers attach various “bogus” texts to the email, together with the attached image, that appears in genuine emails and not in spam. Figure 1 shows the design of spam location process.

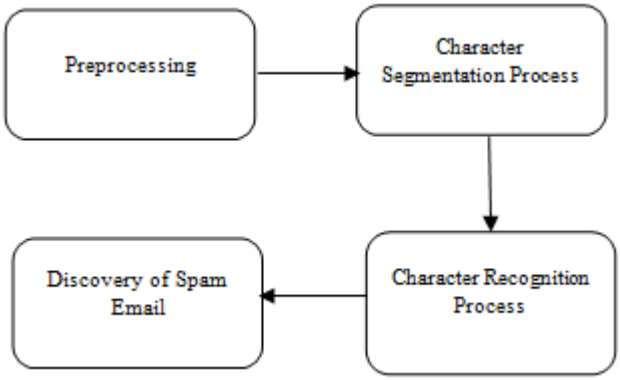


Figure 1 Design of spam detection process

In pre-processing, the colour image is transformed into grey scale image. Grey-scale image conversion to binary format is the method of detaching the noise from black and white pixel images. It eliminates the connected components i.e. objects which are less than 15 pixels from the binary image. For recognizing a character, the skewed angle of a character will be detected and corrected. The character segmentation process is categorized into classifier based and non-classifier based techniques. Existing methods of character segmentation process are discussed and analyzed.

2.1 Character Segmentation

Discrete Wavelet Transform

Character segmentation calculation continuously DSP based tag utilizing 2D Haar Wavelet Transform could be utilized [5]. Enhanced image edges and improved LP district recognition for its appropriateness progressively application are the highlights of the calculation. The Haar WT identifies three kinds of edges utilizing a solitary channel while customary techniques, for example, Sobel would require more than one cover for the activity. DWT is a particular instance of sub-band separating and estimation done utilizing channel bank. The sign is gone through high-pass and low-pass channels simultaneously to create sifted yield. The process of LP detection is edge detection within LP region through grey scale variation, and the Haar edges are compared with grey scale variations to validate the edges. If edges are matched then a rectangle of connecting edges is drawn. Histogram analysis verifies the character extraction and compute bounding box. The experimental results showed an improvement in 2D Haar WT of character segmentation. Results proved that the method could identify the majority edges in image, less noise, and high rate of character segmentation. The challenging factor of character segmentation in license plate is due to rain drops, number plate broken due to accidents, or uneven luminance.

Discrete Wavelet Transform and Gradient method extracted text from images. The input/given image is preprocessed and the Daubechies DWT is applied to acquire 3 types of texture and edges [6]. Daubechies wavelet has high recurrence coefficient range than the Haar wavelet. It decomposes the signal into four different components of frequency domain such as LL, HL, LH, and HH respectively. In high contrast of text region, Gradient difference technique was applied to show the difference from non-text regions. By using Otsu thresholding, non-textual information will be removed. The drawback of the proposed method is when the pixel value is less than global threshold value, it is measured as noise and the text or character region gets eliminated. However, elimination of false positive remains as a challenging task.

Hough Transform

Text segmentation in document image is based on Hough Transform techniques [7]. Image acquisition for document image recognition is digitized through scanner by manual process. Image is pre-processed to convert colour images to grey scale image. Otsu's method is applied to binarise the image and edges are detected. The Hough transform is applied to extract the line and word as a set of connected words and stored as bmp file for performance analysis. For experimentation, 15 English script pages, Bengali, and mixed script pages are used with 812 lines and 7308 words. Through comparative study, the proposed method showed better performance than other existing methods.

Generalized Hough transform has been applied for Arabic printed document segmentation.Voting process gives the Hough transform forcefulness of missing edge points. Segmenting a character by recognition techniques, an indexed dictionary was created for character recognition. Dynamic sliding window technique is used to recognize cursive Arabic characters. The technique is based on recognizing opening and finishing characters of the sub-words, and then middle characters are identified. For each end character put away in the word reference, a similar method is rehashed from left utmost of the starting character to recognize the center character. GHT can be utilized in OCR not exclusively to perceive characters yet additionally to investigate this particular property for the Arabic cursive character without reestablishing in the segmentation arrange. For experimentation, Arabic printed characters of different font and different sizes were used wherein 93% of recognition accuracy was achieved.

Integrated Approach

An integrated approach of License plate detection is proposed [8] using Harris Corner calculation method. A few frameworks were offered for tag acknowledgment, and every technique has specific focal points and confinements. The fundamental advance in automatic license plate recognition framework is the itemized constrainment of vehicle number plate, segmentation, recognition. The segmentation is practiced by a technique for associated part investigation combined with aspect ratio, height of characters and pixel count. The good image and challenged image are taken for experimentation with result of the achievement pace of segmentation exactness got at 93.84%.

2.2 Character Recognition

Recognition of character is the procedure of detecting and recognizing fonts or characters or text from given or input image. Recognition of character is the method of transforming images of printed text, handwritten text, or typewritten into an arrangement implicit by machines for the intention of decline in storing size, editing and indexing/searching [9]. Recognition of character is difficult when document used is of poor quality. Problems occur in recognizing a character when the font size is small.

Character recognition is a challenging task when different font types are used. In the pre-processing phase, the characters are recognized by skew detection and skew correction.

Pre-processing

A filtered archive image utilized [10] skew estimation and revision calculation dependent on Shear let change. It manages heading selectivity and time-recurrence limitation. It was plausible to recognize the skew alignment of the document image exactly. Exploratory outcomes demonstrated that the proposed calculation has a high precision pace of skew estimation in any event, when the filtered archives images contain noise or a few images or graphs.

Scientists [11] applied a skilled skew recognition and amendment technique for Arabic transcribed text/content line dependent on sub-words jumping. The offered technique approximates a text or character line developed on processing the essential issue for its auxiliary words jumping. At that point the text/content line parts on the surveyed standard are adjusted. It was examined with the level projection technique as far as viability. The proposed strategy got an exactness proportion of 96.1 percentage, and 6.7 sec as usual. It can likewise consequently distinguish char/content measure of document whichever direction.

A nonexclusive and strong strategy [12] which can adapt to a broad assortment of document types and hand written frameworks. The disadvantages of the proposed framework are as per the following: when applied to enormous non-text/content zones with no vertical and horizontal lines noisy projection profiles are formed; italic textual style types cause predisposition in the vertical estimation.

A survey [13] on skew detection and correction on document image. The setback was of critical importance in the making libraries digitalization projects, preset contented conversion systems domain. The relationship among the main types of skew detection algorithms presents the merits and demerits, as well as proposed improvements. The drawback of Hough Transform approach is complexity, and error occurs if input page contains images.

OCR-based Character Recognition

This technique extracts and examines the text or character embedded into images, whereas other techniques can be establish in profit-making spam filters [14]. Keyword recognition is a basic technique to confirm the spam mail by checking the event of occurrences of typical keyword that shows up in spam messages.

Text categorization examines whether a similar text classification strategies functional to e-mail's body content can be compelling additionally to investigate the text or character extract using OCR [15]. Text classifiers prepared on content originating from e-mail's body are considered and tried on content originating from both the e-mail's body and joined images (assuming any). This strategy permits the impromptu creation of image spam detection rate by dismissing OCR mistakes. The perception ensnares that the spam image identification pace is better when the character or content removed by optical character recognition is handled by an unmistakable text classifier. The explanation is the text utilized for training and testing is influenced by similar sorts of OCR mistakes. Text categorization procedure isn't examined against obfuscated images and is just implemented in the Bayes OCR Spam Assassin module, which encourages the text extricated by OCR to the content classifier remembered for Spam Assassin.

Text Extraction

A methodology [16] used to extract text from different images The authors have use discrete wavelet transform (DWT) for extract text or content from multifaceted images. Sobel edge detector is used to extract edges of the text. Text extractions from complex images such as texture based approach and region based approach are two different approaches were used.

A vigorous methodology for text or character extraction and detection in images. To begin with, the input image is separated by the Median filter any noises. At that point edges are recognized utilizing LOG edge locator. The morphological expansion activity is applied for object localization. All the connected components are then extricated and all non-content character segments are disposed of by a two-advance step procedure. Features are then extricated from the removed Components. These feature structure from component vector for SVM. These features are tested or tried with SVM for perceiving individual characters. At that point, every single perceived character are converged to frame text lines.

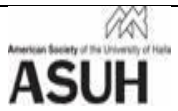
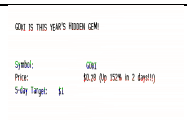
III. RESULT ANALYSIS

The proposed algorithms that will improve the detection accuracy of text and image based email classification are experimented to achieve the research objectives. All the proposed algorithms are fulfilled using MATLAB (version R 2013a), and the experimentations are performed on an Intel(R) Core (TM) i5 machine with a speed 2.60 GHz and 8.0 Giga Bytes RAM. For experimentation, images were taken from image spam dataset. The proposed approaches (Character segmentation using DWT and Hough transforms, Template matching and Contour analysis, and Shape based feature extraction) are evaluated for performance wherein the samples of images would be 23 for testing and 560 images to measure accuracy by F-measure, precision, recall, true negative, true positive, false positive, false negative,

The present research considered the image spam dataset known as Dredze dataset is used for the evaluation of the three proposed tasks: character segmentation, character recognition and spam email detection. As mentioned earlier, the dataset of image spam is collected from [17]. These image spam message are sufficient and robust for the segmentation and recognition evaluation with regard to spam detection domain. The dataset consists of 2173 images in SpamArchive corpus, number ham and spam image are 2359 images, 1248 respectively.

There are ‘wild background’, ‘randomised’, and ‘text only’ images. Text alone present in ‘text only’ images where as randomized images contains random stripes, color shades and color pixels. Table I shows performance metrics used to detect that given images either spam or ham. Figure 2 shows accuracy comparison with the proposed method.

Table I Performance metrics of HAM/ SPAM using proposed approach

No.	Images	Execution metrics
1		THIS IS HAM IMAGE Correct Rate: 85.4839% Error Rate: 14.5161% Accuracy: 96.7742%
2		THIS IS HAM IMAGE Correct Rate: 83.871% Error Rate: 16.129% Accuracy: 93.5484%
3		SPAM IS DETECTED Correct Rate: 82.2581% Error Rate: 17.7419%
4		SPAM IS DETECTED Correct Rate: 80.6452% Error Rate: 19.3548% Accuracy: 93.5484%

Few images are shown here as sample output result in the above table. Number of images is taken from the dredze dataset as sample image and calculated correct rate, error rate based on true positive, true negative, false positive and false negative measures.

Table II delineates results are assessed dependent on the presentation measurements, for example, specificity, sensitivity, F-measure, recall, precision and percentage of accuracy. The Average estimation of execution measurements got for this proposed calculation is about correct rate of 82.2, error rate of 17.7, sensitivity of 86.6, specificity of 81, precision of 0.9, recall of 0.8, F-measure of 0.8 and 95.7 percentage of overall accuracy is calculated. The output shows improvement in terms of specificity and sensitivity respectively.

Table II Output results for SPAM images (n=5)

SPAM							
Correct rate(CR)	Error rate	Sensitivity	specificity	precision	Recall	F-measure	Accuracy
82.3	17.7	100	78	0.909	1	0.95	96.7
85.4	14.5	100	82	0.909	1	0.9523	96.7742
82.26	17.742	83.33	82	0.9091	0.833	0.87	95.16
77.42	22.6	75	78	0.91	0.75	0.822	96.8
83.9	16.13	75	86	0.91	0.75	0.823	93.56

Table III delineates results are assessed based on five various ham images. Normal worth acquired with anticipated algorithm are correct rate of 82.2, error rate of 17.7, sensitivity of 86.6, specificity of 81, precision of 0.9, recall of 0.8, F-measure of 0.8 and overall accuracy of 95.7 percentage are calculated.

Table III Output results for HAM images (n=5)

HAM							
Correct rate(CR)	Error rate	Sensitivity	specificity	precision	Recall	F-measure	Accuracy
82.261	17.72	100	79.1	0.89	1	0.95524	95.161
82.321	17.62	100	78	0.902	1	0.958	95.68
74.2	25.81	100	68	0.91	1	0.95	95.2
82.25	17.75	66.6	86	0.88	0.67	0.75	95.13
80.64	19.35	91.67	78	0.863	0.916	0.9128	93.548

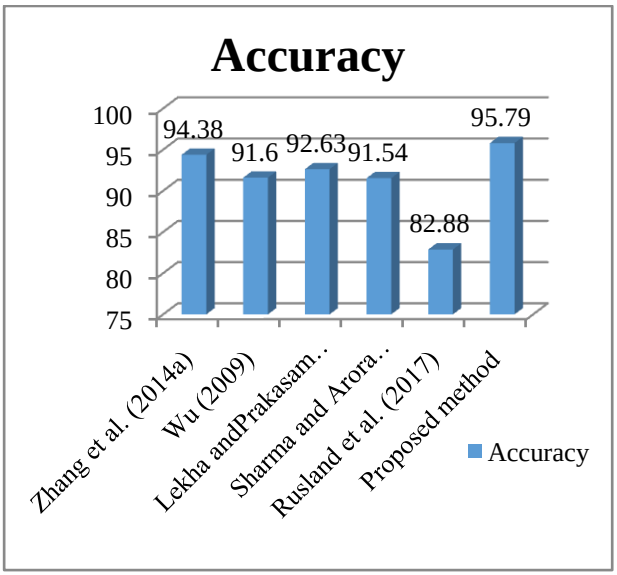
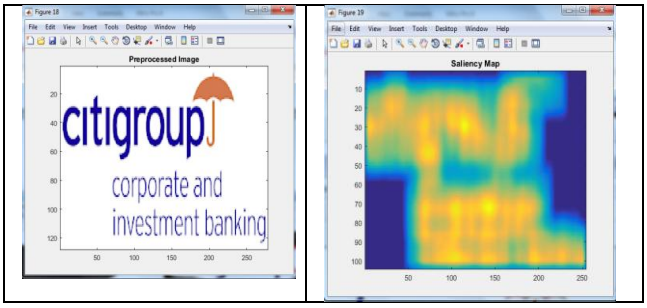


Figure 2 Accuracy comparison

Figure 3 shows screen shot results of image preprocessing, saliency map, mean feature map and character image which is extracted from image. Screen shot shows character extraction process using proposed dredze dataset.



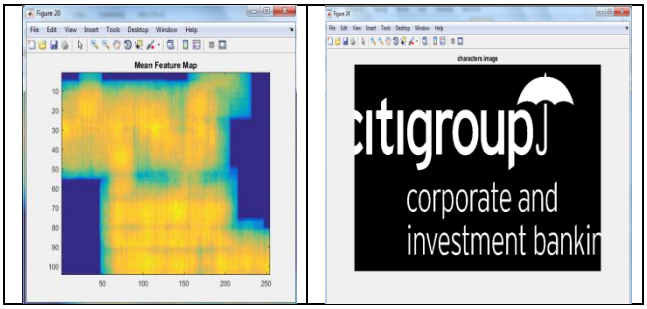


Figure 3 Screen shot of results

IV. CONCLUSION

This paper is to analyze various algorithms, techniques, talk about benchmark statistics/data and execution assessment and to bring up promising headings for future work. The core work is to show how image spam mails are detected using machine learning approaches such as character segmentation and character recognition by overcome the issues of blocked texts. The future enhancement of this work can be performed with other classification techniques to tests the performance analysis.

REFERENCES

1. Radicati, S. & Hoang, Q. (2012). Email Statistics Report. [Online]. PALO ALTO. Available from: http://www.radicati.com/wp/wpcontent/uploads/2012/04/Email-Statistics-Report-2012-2016-Executive-Summary.pdf.
2. Kamboj, R. (2010). A Rule Based Approach for Spam Detection. Thapar University.
3. Statista (2017). Global spam volume as percentage of total e-mail traffic from January 2014 to September 2016, by month. [Online]. 2017. The Statistics Portal. Available from: http://www.statista.com/statistics/420391/spam-email-traffic-share/. [Accessed: 3 January 2017].
4. Biggio, B., Fumera, G., Pillai, I. & Roli, F. (2011). A survey and experimental evaluation of image spam filtering techniques. Pattern Recognition Letters. [Online]. 32 (10). p.pp. 1436–1446. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0167865511000936.
5. Wright, A., Walker, J.P., Robertson, D.E. & Pauwels, V.R.N. (2017). A Comparison of the Discrete Cosine and Wavelet Transforms for Hydrologic Model Input Data Reduction. Hydrology and Earth System Sciences Discussions. [Online]. pp. 1–23. Available from: http://www.hydrol-earth-syst-sci-discuss.net/hess-2017-26/.
6. Syal, N. & Garg, N.K. (2014). Text Extraction in Images Using DWT, Gradient Method And SVM Classifier. International Journal of Emerging Technology and Advanced Engineering. [Online]. 4 (6). pp. 477–481. Available from: https://pdfs.semanticscholar.org/b909/b943041fb372f2b5b865abd73679ae2e7a8b.pdf.
7. Saha, S., Basu, S., Nasipuri, M. & Basu, D.K. (2010). A Hough Transform based Technique for Text Segmentation. Journal Of Computing. [Online]. 2 (2). pp. 135–141. Available from: http://arxiv.org/abs/1002.4048.
8. Panchal, T., Patel, H. & Panchal, A. (2016). License Plate Detection using Harris Corner and Character Segmentation by Integrated Approach from an Image. International Conference on Communication, Computing and Virtualization. [Online]. 79 (1). pp. 419–425. Available from: http://www.academia.edu/24314198/License_Plate_Detection_using_Harris_Corner_and_Character_Segmentation_by_Integrated_Approach_from_an_Image.
9. Kapoor, R., Gupta, S. & Sharma, C.M. (2011). Multi-Font/Size Character Recognition and Document Scanning. International Journal of Computer Applications. [Online]. 23 (1). pp. 21–24. Available from: https://pdfs.semanticscholar.org/7f0f/ed910ec1cb0f2704b2a807335d9c365fd98d.pdf.
10. Lu, Y.M., Sun, B.A., Zhao, L.Z., Wang, W.H., Pan, M.X., Liu, C.T. & Yang, Y. (2016). Shear-banding Induced Indentation Size Effect in Metallic Glasses. Scientific Reports. [Online]. 6. p.p. 28523. Available from: http://www.nature.com/articles/srep28523.
11. Al-Shatnawi, A.M. (2014). A skew detection and correction technique for Arabic script text-line based on subwords bounding. In: 2014 IEEE International Conference on Computational Intelligence and Computing Research. [Online]. December 2014, IEEE, pp. 1–5. Available from: http://ieeexplore.ieee.org/document/7238501/.
12. Stahlberg, F. & Vogel, S. (2015). Detecting dense foreground stripes in Arabic handwriting for accurate baseline positioning. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). [Online]. August 2015, IEEE, pp. 361–365. Available from: http://ieeexplore.ieee.org/document/7333784/.
13. Tanase, M.C., Zaharescu, M. & Bucur, I. (2013). Upsampling-Downsampling Image Reconstruction System", Journal of Information Systems & Operations Management (JISOM). The Proceedings of Journal ISOM. [Online]. 7 (2). pp. 294–299. Available from: http://jisom.rau.ro/downloads/JISOM-10-2-dec-2016.pdf.
14. Samosseiko, D. & Thomas, R. (2006). The Game Goes On: An Analysis of Modern Spam Techniques. In: VB Conference. 2006, sophos.
15. Fumera, G., Pillai, I. & Roli, F. (2006). Spam filtering based on the analysis of text information embedded into images. Journal of Machine Learning Research. [Online]. 7 (1). pp. 2699–2720. Available from: http://www.jmlr.org/papers/volume7/fumera06a/fumera06a.pdf.
16. Gupta, N. & Banga, V.K. (2012). Localization of Text in Complex Images Using Haar Wavelet Transform. International Journal of Innovative Technology and Exploring Engineering. 1 (1). pp. 1–111.
17. Dredze, M., Gevaryahu, R. & Elias-Bachrach, A. (2007). Learning Fast Classifiers for Image Spa. In: proceedings of the Conference on Email and Anti-Spam. 2007, CEAS.
18. Zhang, Y., Wang, S., Phillips, P. & Ji, G. (2014a). Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems. [Online]. 64. pp. 22–31. Available from: http://dx.doi.org/10.1016/j.knosys.2014.03.015.
19. Wu, C.H. (2009). Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert Systems with Applications. [Online]. 36 (3 PART 1). pp. 4321–4330. Available from: http://dx.doi.org/10.1016/j.eswa.2008.03.002.
20. Lekha, K.C. & Prakasam, S. (2016). Prediction of Respondants' Knowledge towards Cyber Security measures using various Classification Algorithms. International Journal of IT and Knowledge Management. [Online]. 10 (1). pp. 1– 5. Available from: http://csjournals.com/IJITKM/PDF10-1/1. Chitra.pdf.
21. Sharma, S. & Arora, A. (2013). Adaptive Approach for Spam Detection. IJCSI International Journal of Computer Science. [Online]. 10 (4). Available from:https://pdfs.semanticscholar.org/956c/dfa8574d01f0cdb2eaa5383ea5028a1eadc6.pdf.
22. Rusland, N.F., Wahid, N., Kasim, S. & Hafit, H. (2017). Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets. IOP Conference Series: Materials Science and Engineering. [Online]. 226. p.p. 12091. Available from: http://stacks.iop.org/1757-899X/226/i=1/a=012091?key=crossref.262a040d608eaab52b8501086da85f26.

AUTHORS PROFILE

Mallikka Rajalingam received her M.Sc Information Technology from Bharathidasan University, Tiruchirappalli, India in 2005, M.Phil Computer Science from Madurai Kamaraj University, Madurai, India in 2008, M.Tech Computer Science & Engineering from SASTRA University, Thanjavur, India in 2009. She worked as a Research Officer (RO) at School of Computer Science, Universiti Sains Malaysia (USM), Malaysia. She is currently pursuing the Ph.D. degree at the Department of Computer Science & Engineering, Bharathidasan University, Trichy, India. Her research interests include image processing, computer vision, pattern recognition, character recognition, document image analysis, text analysis and multimedia networking.

Dr. M. Balamurugan is currently working as Professor and Head in the Department of Computer Science and Engineering of Bharathidasan University, Trichy, India. He has credits of 20+ international and national conferences publications. He has published 30+ research papers in national and international journals. His research interests are mainly focused on the area of Data Science. He has supervised several research scholars in these areas.