Naïve Bayes Filter for Communication & Enhancing Semantic in Email

Mariyan Richard A, Prasad Naik, Suhas A, Drakshaveni G

Abstract: ***Due tothe current pandemic of COVID-19, the world has turned into ONLINE modeand an increase in online communication thereby information exchange, sharing useful data through emails and other social Medias.So addressing the security issues places a vital role in computer security and shouldhave thepriorities. We need a security check to enhance the inbox so that the important information or emails should not reach to the spam box. In this paper to improve the filtering techniques, wehave adopted the Naïve Bayes approach in implementation and enhancing the spam filter in the email. Bayes's approach is efficient, accurate, and simple in implementing the proposed algorithm. Bayes algorithm is used to verify correct semantic information of the email and avoidsthe pass to pass approach if the incoming mail is important. The Python language is used to develop the proposed algorithm.***

Keywords: Naïve Bayes, String Sematic, Spam Filtering, Python Language.

I. INTRODUCTION

Spam E-Mail is an unconstrained and unwanted messages sent in bulk. Normally, spam is sent for business and marketing purposes. It might be sent in immense volume by botnets, frameworks of polluted PCs. While a couple of individuals consider it to be corrupt, various associations despite everything use spam. The cost per E-Mail is incredibly low, and associations can pass on mass sums dependably. Spam E-Mail can moreover be a threatening undertaking to get to your PC. Botnets are an arrangement of as of late defiled PCs. Along these lines, the main spammer can be difficult to follow and stop [1].

E-Mail filtering is the process of blocking its contents based on a pre-defined set of rules. It can apply to the intercession of human knowledge, yet regularly alludes to the programmed handling of approaching messages with hostile to spam procedures - to active messages just as those being gotten. E-Mail separating programming may dismiss a thing at the underlying SMTP association stage or pass it through unaltered for conveyance to the client's letterbox - or on the other hand: divert the message for conveyance somewhere else; isolate it for additional checking; alter or 'tag' it here and there. E-mailbox service providers can incorporate dedicated lines in the transmission as a feature of the entirety of the recipients. Against infection, hostile to spam, URL sifting, and validation based dismissals are basic channel types [2-4].

Revised Manuscript Received on November 20, 2020.

Mariyan Richard A, Assistant Professor, Dept. Of MCA NMIT, Bengaluru, India E-mail Mariyanrich01@gmail.com

Dr. Prasad Naik Hamsavath, HOD,Dept.Of MCA. NMIT, Bengaluru, India. E_mail. naikphd@gmail.com

Suhas A, Dept. Of MCA NMIT, Bengaluru, India E-mail. suhas.a1996@gmail.com

Drakshaveni G, Assistant Professor, Dept. of MCA, BMSIT, Bengaluru, India. E-mail. drakshavenig@bmsit.in

Corporations frequently use channels to ensure their workers and their data innovation resources. A catch-all channel will "get all" of the messages routed to the area that doesn't exist via the post office server - this can help abstain from losing messages because of incorrect spelling. Users might have the option to introduce separate projects or arrange sifting as a feature of their E-Mail program (E-Mail customer). In E-Mail programs, clients can make individual, "manual" channels that at that point consequently channel e-mail as indicated by the picked measures. E-mail channels can work on inbound and outbound E-Mail traffic. Inbound E-Mail sifting includes checking messages from the Internet routed to clients ensured by the separating framework or for legal interference. Outbound E-Mail sifting includes the opposite - filtering E-Mail messages from nearby clients before any possibly destructive messages can be conveyed to others on the Internet [3,6].

Existing E-Mail Spam Filtering frameworks are subject to List-Based Filter procedures, for example, Blacklist, Real-Time Blackhole List, Whitelist, and Greylist. The boycott is a well-known spam-sifting strategy endeavor to stop undesirable E-Mail by filtering messages from the pre-set rundown of senders that your association's framework overseer makes [4,7,8]. Boycotts are records of E-Mail locations or Internet Protocol (IP) addresses that have been recently used to send spam. At the point when an approaching message shows up, the spam channel verifies whether it's IP or E-Mail address is on the boycott; assuming this is the case, the message is viewed as spam and dismissed. Though boycotts guarantee that realized spammers can't arrive at clients' inboxes, they can likewise misidentify authentic senders as spammers. These alleged bogus positives can result if a spammer happens to send garbage e-mail from an IP address that is additionally utilized by genuine E-Mail clients. Additionally, since numerous shrewd spammers routinely switch IP delivers and E-Mail delivers to cover their tracks, a boycott may not promptly get the most current flare-ups.

II. LITERATURE REVIEW

WuxuPeng : Author in his paper he explained about how important is security concerning the online platform and how is Naïve Bayes algorithm has disadvantages like not properly classifying emails when they contain leetspeak or diacritics. So he explains how his proposed work improves a Novel algorithm for enhancing the accuracy of the Bayes algorithm. He used python as a programming language to implement the work and used concepts of semantic-based, keyword-based, and machine learning algorithms to increase the accuracy[15,16,17].

Deepika Mallampati, Nagaratna P. Hegde: Authors explains about the spam emails were have alias name has non-self, unsolicited commercial emails or fraudulent emails sent to a group of people or for a company. He used Machine learning algorithms. Machine learning classifier to check whether the email received is a valid message or an unwanted message. They used Deep learning as potential tactics that can tackle the challenges of spam emails efficiently [10,11,12].

Jon Kågström: The author explains that Witten Bell is good at Turing with small performance loss comparative to simple Good Turing. He also explained Robinson's estimate based on Bayesian smoother showed excellent results and easy to implement and less computationally expensive than both Witten Bell and Good Turing [13,14].

III. SPAM FILTER

The proposed framework receives Content-Based Filters, which instead of upholding no matter how you look at its strategies for while sending Communication messages from any specific E-Mail or IP address, content-based channels assess words or expressions found in any individual message to decide if communication is spam or not spam [8].The E-mail Spam filter has previously been based on fetching spam signature via supervised learning using communication messages through emails explicitly manually labeled as spam or not spam. In this paper, we study of unsupervised machine learning based spam filter for more effectively identify new spamming. Communication spam filter identifies the unsolicited, unwanted, and virus-infested email as we call them to spam emails and stop it from getting into email inboxes. Internet service providers use spam filters to make sure they aren't distributing spam. The best spam filters currently available in the market are 1)Spam Titan(TitanHQ) it is suitable for all types of businesses2) ZERO SPAM 3) Spambrella 4)MailChannels5) Xeams 6) Topsec Email security 7) Symantec email security 8) MailWasher [18,19,20]

Since a Naïve Bayes channel is continually fabricating its assertion lists dependent on the messages that an individual client gets, it hypothetically turns out to be increasingly powerful the more it's utilized. In any case, since this technique requires a preparation period before it begins functioning admirably, you should practice persistence and will presumably need to physically erase a couple of garbage messages, at any rate from the start.[21,22]

IV. METHODOLOGY

Mathematical Model for SPAM Filter

1. We have to compute the probability that the message is spam, knowing that a given string appears in the message.
2. Then we compute the probability that the message is spam, taking into consideration all of its words
3. Then finally we give with rare string

To minimize false positives and increase the accuracy of Naive Bayes, an addition to the existing Naive Bayes method was created. This addition will be able to convert symbols inside words to possible letters and use a spell check function to ensure the corrected symbol is a word and then run the word through the Naive Bayes spam filter [23,24]

A. Naive Bayes Classifier -

Naïve Bayes algorithm is a Basic, Statistical technique for handling e-mail filtering for Naive Bayes

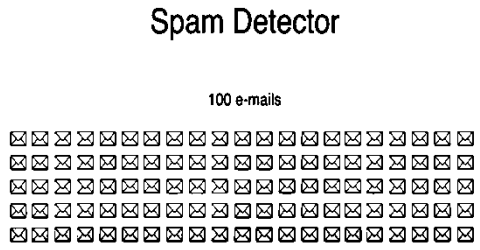
B. Bayesian Classifier – Spam filtering /detection

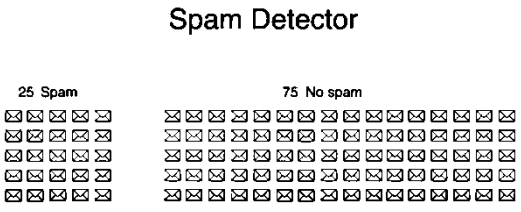
S: Spam
!S: Ham/not Spam

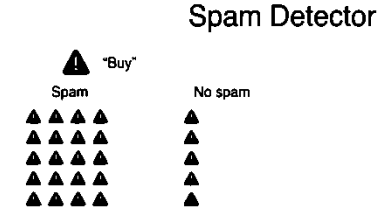
C. Multinomial Naive Bayes – The multinomial Naïve Bayes Classifier is suitable for classification with discrete features like word count in text classification. The multinomial distribution normally requires integer feature count, however, in practice fractional counts such as tf-idf may also work [3]. Theoretically, the best class is determined by multiplying all the probabilities that each word is spam together as shown in the equation to get an overall probability, with probabilities closer to 1 being spam. However, there are instances where the spam word does not occur at all in a message; Laplacian Smoothing may ameliorate this problem.[25]

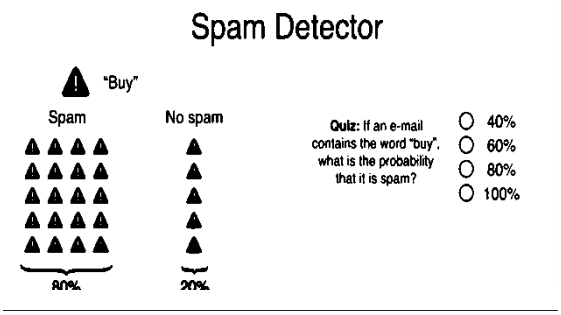
D. Proposed Algorithm – The algorithm is implemented using python language, her in developing the code we used key-based, word-based,semantic-based in unsupervised artificial machine learning [12]

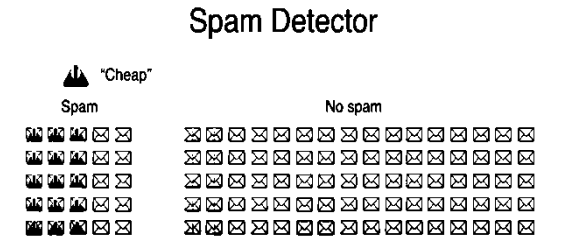
$$P\left(\frac{S}{W}\right) = P\left(\frac{W}{S}\right)P(S)/P\left(\frac{W}{S}\right)P(S) + P\left(\frac{W}{!S}\right)P(!S)$$

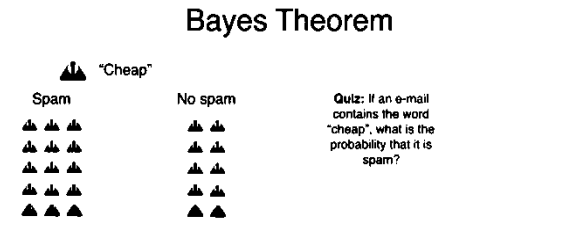


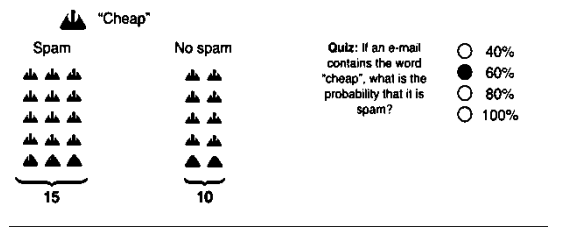


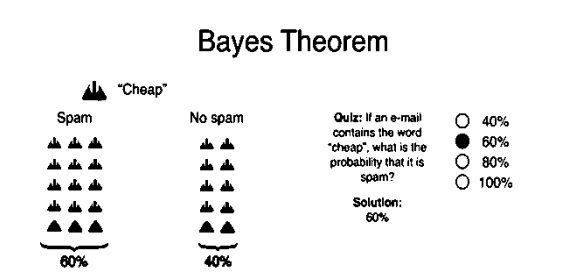


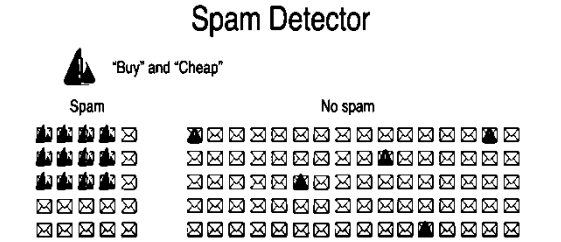


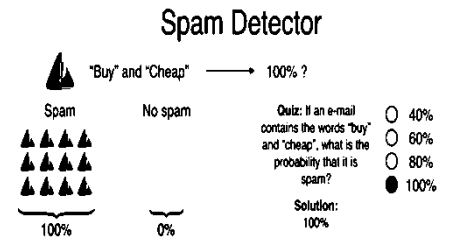


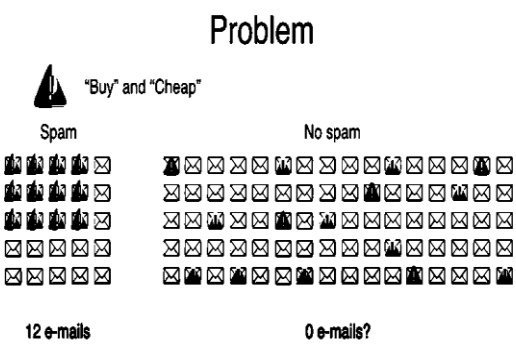












If an emailcontains the word" BUY" what is the probability that it is a SPAM
40%,60%,80%,100%

	Total	BUY	Cheap
Data set 1:	100	5	10
Data set 2:	25	10	15
Data set 3:	75	5	10

$$P\left(\frac{S}{B}\right) = P\left(\frac{B}{S}\right)P(S)/(P\left(\frac{B}{S}\right) + P\left(\frac{B}{H}\right)P(B)$$

$$P(spam\ if\ \text{BUY}) = \frac{20}{25} * \frac{\frac{25}{100}}{\frac{20}{25} * \frac{25}{100} + \frac{5}{75} * 75/100}$$

$$P(spam\ if\ \text{BUY\&"Cheap"}) = \left(\frac{20}{25} * \frac{15}{25} * \frac{25}{100}\right) \Big/ \left(\frac{20}{25} * \frac{15}{25} * \frac{25}{100} + \frac{5}{75} * \frac{10}{75} * \frac{75}{100}\right)$$

$$= 94.737\%$$

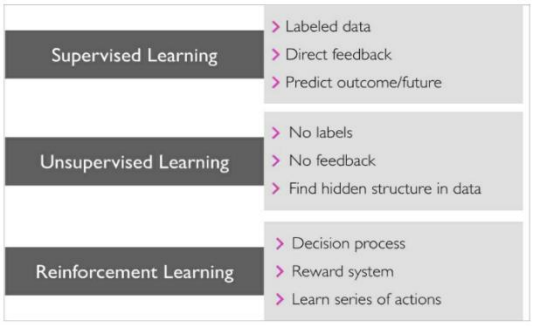


Fig. 1. Making predictions with supervised learning.

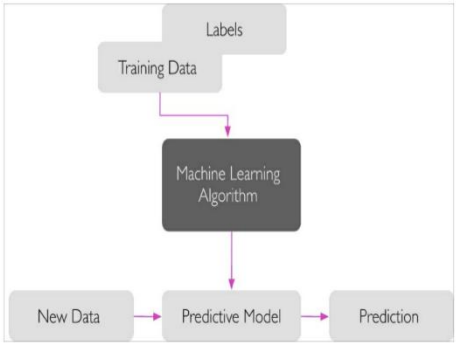


Fig. 2. Proposed algorithm flow.

Consider the data sets

Data sets	String	Filter
1	I Loved the Movie	Ham
2	I hated the Movie	Spam
3	A Great Movie, Good Movie	Ham
4	Poor acting	Spam
5	Great Acting, A Good Movie	Ham

There are 10 unique words in this data set
They are
I, Loved, the, Movie, hated, a ,Great, Poor, Acting , Good

$P(ham) = \frac{3}{5} = 0.6;$

Data set	**I**	**Loved**	**the**	**movie**	**hated**	**a**	**great**	**poor**	**acting**	**good**	**Class**
1	1	1	1	1							Ham
2	1		1	1	1						Spam
3				2		1	1			1	Ham
4								1	1		Spam
5				1		1	1		1	1	Ham

$P\left(\frac{I}{ham}\right) = \frac{1+1}{14+10} = 0.0833$	$P\left(\frac{loved}{ham}\right) = \frac{1+1}{14+10} = 0.0833$
$P\left(\frac{the}{ham}\right) = \frac{1+1}{14+10} = 0.0833$	$P\left(\frac{movie}{ham}\right) = \frac{5+1}{14+10} = 0.2083$
$P\left(\frac{a}{ham}\right) = \frac{2+1}{14+10} = 0.125$	$P\left(\frac{great}{ham}\right) = \frac{2+1}{14+10} = 0.125$
$P\left(\frac{acting}{ham}\right) = \frac{1+1}{14+10} = 0.0833$	$P\left(\frac{good}{ham}\right) = \frac{2+1}{14+10} = 0.125$
$P\left(\frac{hated}{ham}\right) = \frac{0+1}{14+10} = 0.0417$	$P\left(\frac{poor}{ham}\right) = \frac{0+1}{14+10} = 0.0.417$

"I hated the poor acting"

If V_j=ham ;P(ham)P(I/ham)P(hated/ham)P(the/ham)P(poor/ham)P(acting/ham)

$=6.03*10^{-7}$

If V_j=spam ;P(spam)P(I/spam)P(hated/spam)P(the/spam)P(poor/spam)P(acting/spam)

$=1.22*10^{-5}$

$$P\left(\frac{the}{ham}\right) = \frac{1+1}{14+10} = 0.0833$$

$P\left(\frac{I}{spam}\right) = \frac{1+1}{6+10} = 0.125$	$P\left(\frac{the}{spam}\right) = \frac{1+1}{6+10} = 0.125$
$P\left(\frac{movie}{spam}\right) = \frac{1+1}{6+10} = 0.125$	$P\left(\frac{hated}{spam}\right) = \frac{1+1}{6+10} = 0.125$
$P\left(\frac{poor}{spam}\right) = \frac{1+1}{6+10} = 0.125$	$P\left(\frac{acting}{spam}\right) = \frac{1+1}{6+10} = 0.125$
$P\left(\frac{loved}{spam}\right) = \frac{1+1}{6+10} = 0.0625$	$P\left(\frac{a}{spam}\right) = \frac{0+1}{6+10} = 0.0625$
$P\left(\frac{great}{spam}\right) = \frac{0+1}{6+10} = 0.0625$	$P\left(\frac{good}{spam}\right) = \frac{0+1}{6+10} = 0.0625$

$$P(spam) = \frac{2}{5} = 0.4;$$

V. RESULTS

	Spam		Not Spam	
Total	25		75	
BUY	20	4/5	5	1/15
Cheap	15	3/5	10	2/15
Buy & Cheap	12	12/25	2/3	2/225
	$\frac{12}{12+\frac{2}{3}} = \frac{36}{38} = 94.734\%$			

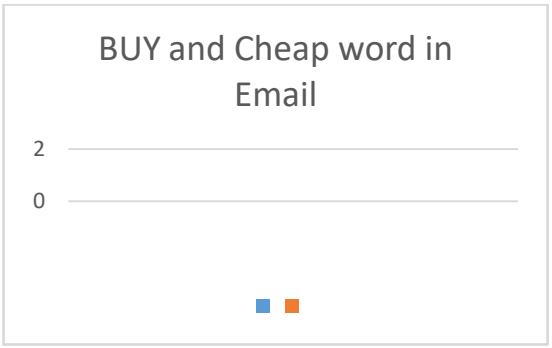


Fig. 3. SPAM Message Detection using the proposed algorithm.

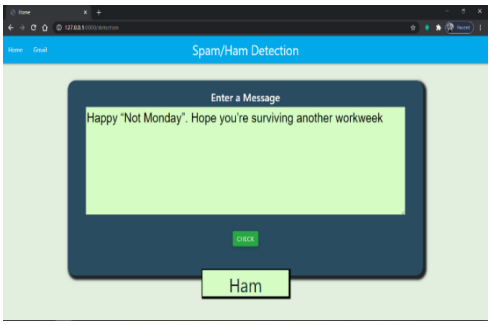


Fig. 4. Ham Message Detection using the proposed algorithm.

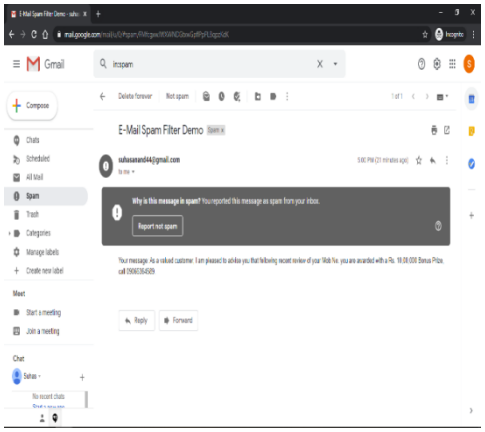


Fig. 5. An E-Mail with SPAM Content being detected by Gmail.

VI. CONCLUSION

Today's generation everyone is using online mode for communication, and started using smart phones, laptops and store/exchange lot of information via emails. We get gigabits of messages every day and partitioning them with spam or not is very difficult talk in hand. So we have come up with new idea of Naïve Bayes spam filter algorithm.

ACKNOWELDEGEMENT

We are extremely gratitude to all reference we used in this paper in providing with best facilities of Naïve Bayes algorithm from different sources like different authors from different papers, Wikipedia, YouTube, etc available material in directly helped in this paper

REFERENCES

1. W. Peng, L. Huang, J. Jia, and E. Ingram, *"Enhancing the Naive Bayes Spam Filter through Intelligent Text Modification Detection,"* 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), New York, NY, 2018, pp. 849-854, DOI: 10.1109/TrustCom/BigDataSE.2018.00122.
2. K. Netti and Y. Radhika, *"A novel method for minimizing loss of accuracy in Naive Bayes classifier,"* 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, 2015, pp. 1-4, DOI: 10.1109/ICCIC.2015.7435801.
3. L. Li and C. Li, "*Research and Improvement of a Spam Filter Based on Naive Bayes,"* 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics, Hangzhou, 2015, pp. 361-364, DOI: 10.1109/IHMSC.2015.208.
4. Guo-Qiang, *"An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification,"* 2010 Second International Conference on Computer Research and Development, Kuala Lumpur, 2010, pp. 699-701, DOI: 10.1109/ICCRD.2010.160.
5. Metsis, Vangelis & Androutsopoulos, Ion &Paliouras, Georgios. (2006). *'Spam Filtering with Naive Bayes - Which Naive Bayes'*. In CEAS.
6. Harryzhang,. (2011). *Exploring Conditions For The Optimality Of Naïve Bayes.*International Journal of Pattern Recognition and Artificial Intelligence. 19. 10.1142/S0218001405003983.
7. Rennie, Jason & Shih, Lawrence & Teevan, Jaime & Karger, David. (2003). *Tackling the Poor Assumptions of Naive Bayes Text Classifiers*. Proceedings of the Twentieth International Conference on Machine Learning. 41.
8. "A bayesian approach to filtering junk e-mail,"https://robotics.stanford.edu/users/sahami/papers-dir/spam.pdf, StanfordUniversity, accessed: 2017, 2018.
9. A. Bhowmick and S. Hazarika, *"Machine learning for e-mail spamfiltering: review, techniques and trends,"* https://arxiv.org/abs/1606.0104,2016, accessed: 2017.
10. "The shifting tactics of spammers: What you need to knowabout new email threats," https://secureemailplus.com/wp/WP10-01-0406 Postini Connections.pdf, Postini, Inc, 2004, accessed: 2017, 2018.
11. "The shifting tactics of spammers: What you need to know about newemail threats," https://emailmarketing.comm100.com/email-marketingebook/spam-words.aspx, accessed: 2018
12. "An analyst review of Hotmail anti-spam technology,"https://www.lifewire.com, Radicati Group, Inc, 2010, accessed:2017.
13. "Exim internet mailer," https://www.exim.org, accessed: 2017, 2018.
14. "Spamassassin: Welcome to SpamAssassin,"https://spamassassin.apache.org, accessed: 2017, 2018.
15. csmining.org, "Ling-spam datasets - csmining group,"http://csmining.org/index.php/ling-spam-datasets.html, accessed: 2017,2018.
16. H. Tschabitscher, "How many emails are sent every day?"https://www.lifewire.com/how-many-emails-are-sent-every-day-117121,2017 accessed: 2017.
17. K. Tretyakov, *"Machine learning techniques in spam filtering,"* in DataMining Problem-Oriented Seminar, 2004.
18. A. Aski and N. Sourti, *"Proposed efficient algorithm to filter spamusing machine,"* in Pacific Science Review A: Natural Science andEngineering, vol. 18, 2016, pp. 145–149.
19. J. Rao and D. Reiley, *"The economics of spam,"* in Journal of EconomicPerspectives, vol. 26, no. 3, 2012.
20. Graham-Cumming, "Does bayesian poisoning exist?"https://www.virusbulletin.com/virusbulletin/2006/02/does-bayesianpoisoning-exist/, 2006.
21. B. Biggio, P. Laskov, and B. Nelson, *"Poisoning attacks against supportvector machines,"* in Proc. of the 29th International Conference onMachine Learning, 2012, pp. 1467–1474.
22. J. Eberhardt, *"Bayesian spam detection,"* at University of MinnesotaMorris Digital Well, vol. 2, no. 1, 2015.
23. K. Asa and L. Eikvil, "Text categorization: a survey,"https://www.nr.no/ evil/to survey.pdf, 1999.
24. M. Sprengers, *"The effects of different bayesian poison methods on thequality of the bayesian spam filter,"* in B.S. thesis, Radboud UniversityNijmegen, Nijmegen, Netherlands, 2009.
25. S. Raschka, *"Naive Bayes and text classification i: introduction and theory,"* https://arxiv.org/abs/1410.5329, Cornell University

AUTHOR PROFILE

Mr. MARIYAN RICHARD A is currently working as an Assistant Professor for the Department of Masters of Computer Applications (MCA) at Nitte Meenakshi Institute of Technology, Yelahanka, Bangalore. He has a total of 12 years of experience in teaching and 1 year of industry experience. He is currently pursuing his Ph.D. from VTU, Belagavi, Karnataka. Mr. Mariyan Richard A pursued his Post Graduate Degree in MCA from Sacred Heart College, Vellore, Tamil Nadu in 2008 and Graduate Degree in BCA from Islamiah College, Vellore, Tamil Naduin2005. His areas of interest for research include Image processing and Mobile Networks. He is also serving as a Board of Examiners coordinator for the Department of MCA at NMIT, Bangalore. His teaching expertise includes Network Simulator, .NET Programming, Cryptography and Network Security, Mobile Application, Computer Networks, DBMS, Unix, C++ and C. He has guided several research scholars and students in their research work and academic projects.

Dr. Prasad Naik Hamsavath is currently working as a Professor and Head of the Department of Master of Computer Applications (MCA) at Nitte Meenakshi Institute of Technology, Yelahanka, Bangalore. He his PhD from Jawaharlal Nehru University (JNU), New Delhi, India. Dr. Prasad N H has more than 15 years of experience in different Government and Private Organizations. He has served as Assistant Director in Software Technology Parks of India (STPI), Ministry of Communication and Information Technology (MC&IT), New Delhi, Government of India and received the Best Employee award by then Director-STPI-Noida. He also served as 'Software Engineer' at Calance Software Pvt Ltd, Gurgaon, Haryana and later he served in Educational Consultants India Limited (EdCIL), Ministry of Human Resource and Development (MHRD), New Delhi, Govt. of India He has authored 03 books and more than 15 books are edited by him on different emerging areas. His major research areas are Mobile Ad-hoc wireless networks, Information Systems, Computer Networks, Cloud computing, Internet of Things (IoT). He has received a prestigious award "Dr. Abdul Kalam Life Time Achievement Award" in the field of Teaching/Training/Research/Administration and also received "Young Faculty" award at 2nd Academic Brilliance Awards from Education Expo TV, New Delhi. He has lifetime membership of ISTE, MCSI and ACEEE. He is also the Program Chair and Chief Editor for the International conference on "Emerging Research in Computing, Information, Communication and Applications" – ERCICA -2013, 2014 , 2015, 2016, 2018 and 2020 editions held at NMIT, Bangalore, India. The books of the Conference were published with Springer and indexed in all top ranked indexing databases. **ERCICA is one of the prestigious events of NMIT and ERCICA is one of the top ranked International Conferences in the country, and is 'Trademarked' by the Controller General of Patents, Designs and Trademarks (CGPDTM), Chennai, Ministry of Commerce and Industry, Govt. of India.** He has visited number of countries including Nepal, Bhutan, Sri Lanka, Bangladesh, Thailand, Mauritius, Ethiopia, Nigeria, Oman, Dubai under Ministry of Human Resource and Development, Govt. of India for the promotion of Indian Education abroad (Study in India Programme) and promotion of NMIT foreign admissions.

Mr. Suhas A is currently working as a Systems Engineer Trainee at Infosys Limited, Mysore, Karnataka. He pursued his Post graduate degree in Masters of Computer Application from Nitte Meenakshi Institute of Technology, Yelahanka, Bangalore in 2020 and his Graduate degree in Bachelors of Computer Applications from Ramaiah College of Arts, Science and Commerce, Bangalore in 2018. He was the Student body president for the Department Association – SPARKS at NMIT, Bangalore. He has coordinated for several events held at both collegiate and intercollegiate levels. Mr. Suhas has worked on several academic projects which include a web Application to automate the college admission and student management. His most recent project involved a Hardware project titled – Automatic Smart Parking System using IoT. He was awarded the 3rd Rank in for his academic performance in MCA at NMIT, Bangalore. Mr. Suhas has also been a South-zone runner up in the WorldSkills – 2015 in the Web Development category, organized by the Government of India.

Drakshaveni G is currently working as a Assistant Professor in the Department of Master of Computer Applications (MCA) at B M S Institute of Technology, Yelahanka, Bangalore. she is pursuing her PhD from Visvesvaraya Technological University (VTU), Belagavi, India. Drakshaveni G has more than 17 years of experience in Teaching she has published papers in both National ,International journals . Her major research areas are Medical Image Processing, Data Base Management, Software testing . she has received a Best paper Award for the paper titled "Embedding data in JPEG & BMP image using LSB & cryptography algorithm" in International conference Global Paradigm shifts AVANT-GRADE-2014 Organized by Bangalore university Teacher's council of commerce & management and Seshadripuram first grade college ,Bangalore , 19th March 2014, and also received best paper award for the paper titled "Color Segmentation using Digital image processing" national conference on "Computing for community services" organized by 27th CSI Karnataka students convention" & Reva Institute of technology & management, Bangalore 27th and 28th March 2013, and also received best projects awards for the project titled "Information Hidding using steganography system approach" has received Best Project conducted by Dept. of MCA, BMSIT June 2014 Received FIRST PLACE in Project contest, Project titled "A Utility for scraping and parsing data in webpages" has received Best Project conducted by Dept. of MCA, BMSIT 18th June 2015 Received SECOND PLACE in Project contest and project titled "Multifunction Smart-Bot Using Arduino" have received Best Project on OPEN DAY conducted by Dept. of MCA ,BMSIT 2019 FIRST PLACE and received APPRECIATION FOR ACHIEVING 100% RESULTS IN THE SUBJECTS HANDLED II Sem MCA – Data structure using C – July 2011 ,II Sem MCA – Data structure using C – July 2012 ,III Sem MCA – Data Base Management system – Jan 2011 ,III Sem MCA – Data Base Management system – Jan 2013 ,V Sem MCA – Software Testing – Jan 2011 ,IV Sem MCA- software Testing – August 2019,V Sem MCA –Dot net Programming- Jan 2019