Asian Journal of Advances in Research

Volume 6, Issue 1, Page 167-179, 2023; Article no.AJOAIR.2465

Machine Learning for SPAM Detection

Phani Teja Nallamothu [a*] and Mohd Shais Khan [b]

[a] *Strava, United States.*
[b] *Osmania University, Hyderabad, Telangana, India.*

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Review Article

Received: 09/01/2023
Accepted: 15/03/2023
Published: 17/03/2023

ABSTRACT

In practically every industry today, from business to education, emails/messages are used. Ham and spam are the two subcategories of emails/messages. Email or message spam, often known as junk email or unwelcome email, is a kind of message that can be used to hurt any user by sapping their time and computing resources and stealing important data. Spam messages volume is rising quickly day by day. Today's email and IoT service providers face huge and massive challenges with spam identification and filtration. Spam filtering is one of the most important and well-known methods among all the methods created for identifying and preventing spam. This has been accomplished using a number of machine learning and deep learning techniques, including Naive Bayes, decision trees, neural networks, and random forests. By categorizing them into useful groups, this study surveys the machine learning methods used for spam filtering. Based on accuracy, precision, recall, etc., a thorough comparison of different methods is also made.

Keywords: Spam; ham; machine learning; supervised machine learning.

1. INTRODUCTION

These days, short message service is a very popular method of communication. This system has millions of users linked to it because of its quick responses, accessibility, and affordable costs. There are two different types of SMS [1]. The first is spam, which counts the number of unsolicited commercial messages a user has received. With these notifications, the user encounters a number of issues, including a slow device and storage concerns [2]. Further, deleting spam from memory takes a long time. Various techniques, such as blacklist, naive bluesman, and keyword matching algorithms, are utilized to identify this spam issue [3-5].

Spam communications have a negative effect on text and email messages today and annoy SMS users. Cybercriminals and various advertising agencies utilize these kinds of spam [6].

**Corresponding author: Email: Phani.teja89@gmail.com;*

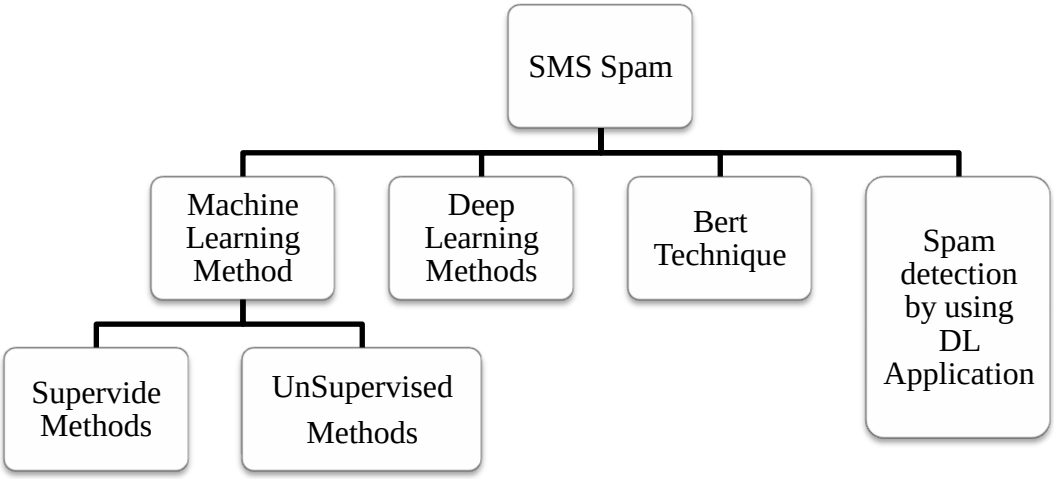

Fig. 1. Machine learning techniques

The fundamental problem with spam is that there is no longer any privacy; when someone responds to these SMS messages, privacy is violated. Because it uses a single click to assault the privacy bridge [7,8]. The easiest tool for a cyber-attack is a mobile phone. Research has shown that more than 200 million mobile users receive spam SMS in a single day, which is insufficient [9-11].

1.1 What Is Spam?

Unwanted and unpleasant text messages in the form of spam are those that we repeatedly getviaa transmission channel. Spam messages have an impact on a device's performance, power, and storage system. In short, spam has proven to be the most unpleasant aspect of personal communication [12,13].

1.2 What Is Ham?

Ham refers to messages that we receive from end devices that are not spam and are on a good list of requested and wanted messages. About 2001, Spam Bayes first used the term "ham," which is currently recognized to mean "e-mail and messages that are commonly appreciated and aren't deemed spam [14,15].

Its utilization is especially normal among anti-spam software developers, and not broadly known somewhere else; as a general rule, it is most likely better to utilize the expression "non-spam", all things considered [16,17].

2. SPAM FILTRATION TECHNIQUES

Spam emails are becoming more and more prevalent in politics, education, chain messaging, stock market recommendations, and marketing [18]. For effective spam identification and filtering, numerous businesses are currently developing various methods and algorithms. In order to comprehend the filtering process, we discuss a few filtering mechanisms in this part.

2.1 The Common Spam Filtering Technique

A filtering system that employs a set of rules and uses those set of protocols as a classifier is known as standard spam filtering. The first phase is the implementation of content filters, which identify spam using artificial intelligence methods. The second phase involves the implementation of the email header filter, which extracts the header data from the email. After that, backlist filters are applied to the emails to weed out spam emails by securing the emails originating from the backlist file. The next step is the implementation of rule-based filters, which identify the sender based on the subject line and user-defined characteristics. Finally, a technique that enables the account holder to send messages is implemented to use allowance and task filters [19-25].

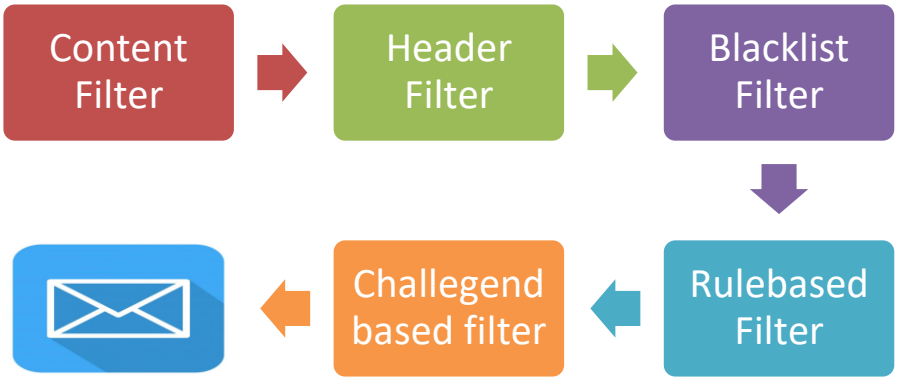

Fig. 2. Approaches to filter spam

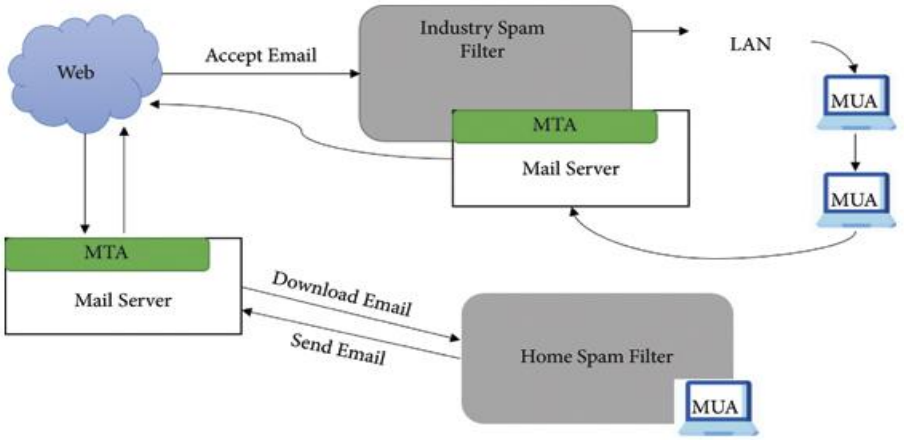

Fig. 3. Client based and enterprise based filtering [31]

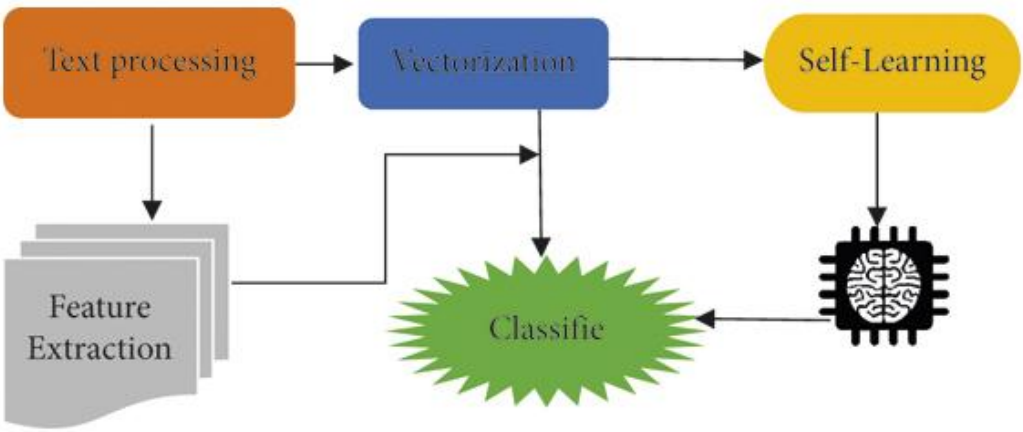

Fig. 4. Case based filtering [37]

2.2 Filtering of Spam on the Client Side

A client is a person who has access to an email network or the Internet and can send or receive emails. Several rules and procedures for ensuring secure communications transmission between persons and organizations are offered by spam detection at the client point. A client needs install various working frameworks on his or her system for data transmission. By connecting with client mail agents and composing, receiving, and handling the incoming emails, such systems filter the client's mailbox [26-28].

2.3 Commercial-Grade Spam Filtering

The process of detecting email spam at the enterprise level involves installing different filtering frameworks on the server, interacting with the mail transfer agent, and categorizing the gathered emails as either spam or ham. This system client employs the system regularly and successfully on a network where emails are filtered using an enterprise filtering technique. The rule of ranking the email is used by existing spam detection techniques. This principle specifies a ranking function and generates a score for each post. A certain score or rating is assigned to the spam or ham message. Since spammers employ various strategies, all jobs are routinely adjusted by adding a list-based technique to automatically block the messages [29-31].

2.4 Spam Filtering Using Cases

The case-based or sample-based spam filtering system is one of the well-known and traditional machine learning techniques for spam detection. With the help of the collection method, this type of filtering has multiple stages; the first one involves gathering data (email). The key change then continues with the client's graphical user interface preparation processes, outlining abstraction, and selection of email data categorization, evaluating the entire process using vector expression and categorizing the data into two groups: spam and genuine email [32-36].

3. MACHINE LEARNING ALGORITHMS

Machine learning methods support data prediction and data classification which are linked to Artificial Intelligence. There are two parts of machine learning methods

- Supervised
- Unsupervised

3.1 Supervised Learning Algorithm

With the help of two datasets, various characterization techniques for SMS spam discovery are evaluated (which were gathered from free sources). Datasets are organized using preprocessing techniques such as tokenization and Tf-IDF, as well as correlation algorithms and deep learning classifiers (the choice tree, SVM, the calculated relapse, ANN, the arbitrary timberland, the AdaBoost, CNN, NB) [38-40].

Three computations, including SVM, NB, and the highest entropy calculation, were used by [6] to identify spam and ham communications. Due to the fact that many of the words are often used, spam is difficult to identify. Fundamentally, SMS spamming is a form of email spamming. It was discovered that SVM provides more accurate results than Gullible Bayes and greatest entropy by using the Spam SMS dataset, which has about 5574 records and is prepared by using stop word evacuation and tokenization. They achieved 97.4% precision using SVM [14,15,38,41,42].

3.2 Unsupervised Learning Algorithm

For spam identification, Weka and RapidMiner, two different arranging tools, were used. They use AI calculations for grouping and ordering, and to verify the precision of these calculations, they used a dataset that can be downloaded from UC Irvine. Findings demonstrate that Weka SVM acquired greater precision of 99.3% in 1.54 seconds for grouping and K-Means in 2.7 seconds is amazing for bunching. With RapidMiner SVM, results are provided in 21 seconds with 96.64% precision and in 37.0 seconds with K-Means [51-56].

3.3 Deep Learning Methods

Here, spam and non-spam messages from clients are separated using convolutional neural organization. The class of spam SMS was identified using Tiago's dataset for its evaluation. To increase the exactness rate, steps for tokenization and stop word preparation are also explained [61-63].

Table 1. Some supervised machine learning methods [44-50]

Methods	Accurate Results	Results	Datasets	Preprocessing Techniques
The decision tree, SVM, NB The logistic regression The AdaBoost ANN CNN The random forest	CNN	99.19%, 98.25%	Spam SMS Dataset (2011-12)	Tf-IDF, Tokenizer
SVM NB MEA	SVM	97.4%	Previously used datasets	Stop words removal, Tokenization
Content-Based techniques (SVM, Clusters techniques)		99.8%	Collected dataset publicly	Tokenization, Stop word removal
Random forest tree, SVM	Boosted SVM	99.14%	Previous dataset	Feature extraction and feature classification
Random forest, SVM, Self-designed feature mapping, DT, logistic regression	Random forest	Precision rate 62.22%	CDR message collection dataset	Tokenization, Stop word removal
KNN, Decision tree-based, Random Forest, CART algorithm, Naïve Bayes, ID3, C4.5, Adaboost algorithm	Random Forest	Random forest accuracy97.2% and without features selection 97.5%	Dataset from the UCI	Tokenization, stop word removal, stemming, Feature extraction, chi-square attribute selection technique
TF-IDF, Random Forest	Random Forest	97.50%	UCI Dataset	Stop word removal, Punctuation correction
KNN1, KNN45, NB, SVM, ITC	KNN1	98.82%	Collected data publicly	Words frequency, Tokenization
SVM, NB, RF, C4.5, Adaboost C4.5, LR, Bagging C4.5, Rough set	Rand Forest	84.40%	Previously dataset	Tokenization, Stop word removal

Table 2. Some unsupervised machine learning methods [58-60]

Methods	Accurate Results	Results	Datasets	Preprocessing Techniques
Content-based (NB, Decision Tree, Logistic Regression, KNN) Non-content- based features	Logistic regression algorithm	97.5%	Indonesian language dataset	Text Normalization, stop word Removal, Stemming, Tokenization
Classification, Clustering	For Classification, SVM is best, For clustering, the K-Means algorithm is best	With Weka SVM 99.3% in 1.54 sec With K-Means 2.7 sec time taken RapidMiner SVM gives 96.64% accuracy in 21 seconds and K-Means gives results in 37.0 seconds.	Downloaded from UCI	Tokenization, Stop word removal

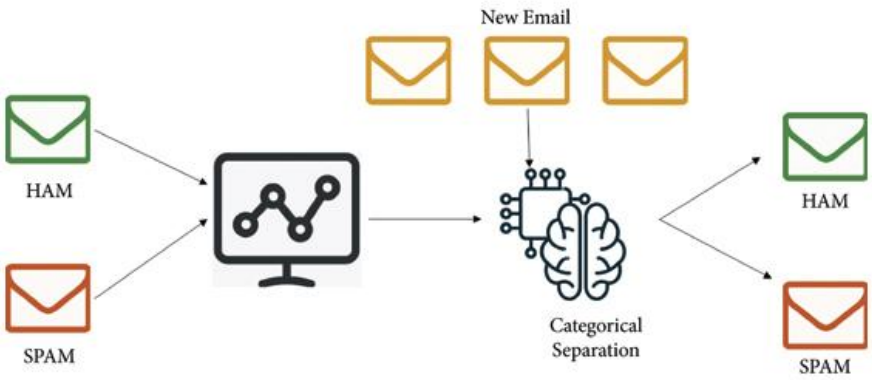

Fig. 5. Supervised machine learning process [43]

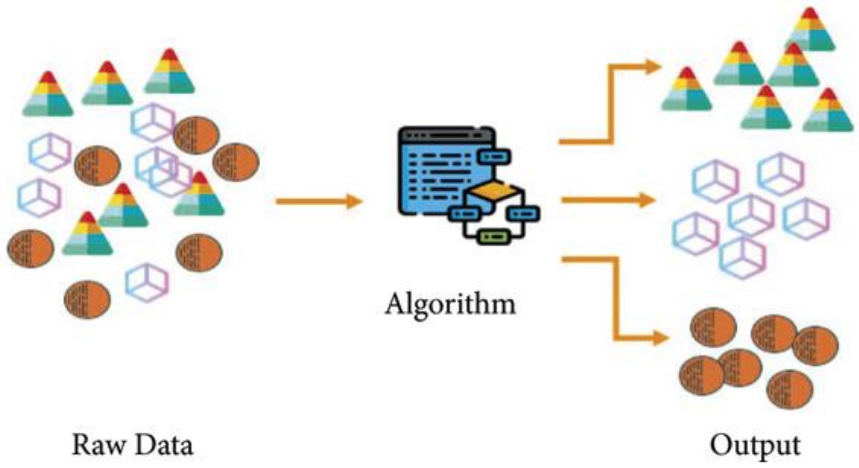

Fig. 6. Unsupervised machine learning process [57]

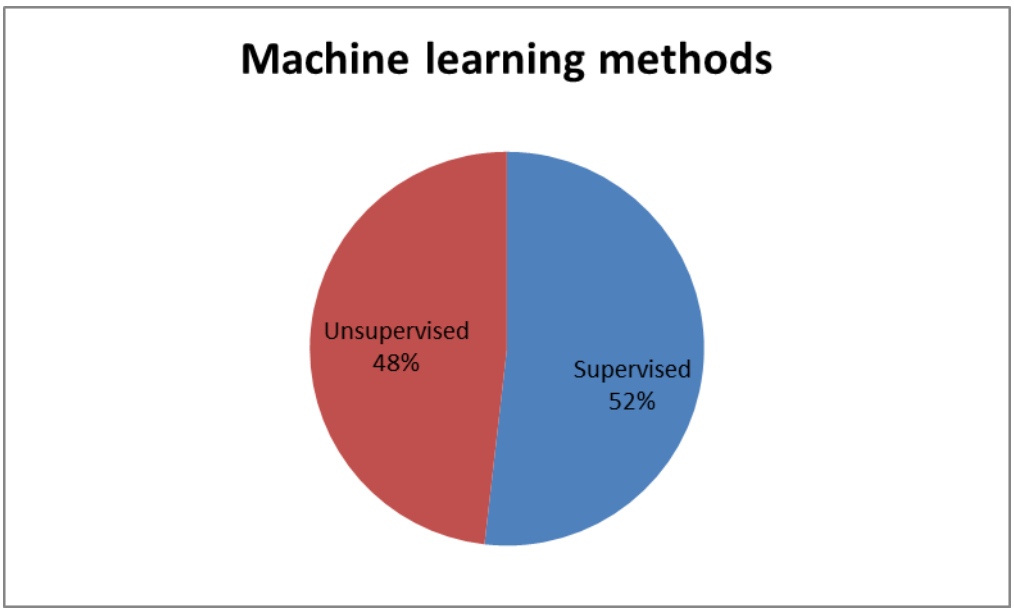

Fig. 7. Supervised vs unsupervised machine learning

Table 3. Some Deep Learning Methods [65-74]

Methods	Accurate Results	Results	Datasets	Preprocessing Techniques
CNN	CNN	98.4%	Tiago	Capitalization, Tokenization, Stop word removal, sentiment analysis, TF-IDF
Evolutionary (NB, K Nearest Neighbor, Decision trees, JRip, CSVM) Non-evolutionary classifiers (Fuzzy AdaBoost, GAssist-ADI, XCS, UCS)	supervised Classifier System (UCS)	93% accuracy with 0% false alarm rate	Real-world dataset	Tokenization, Stop word removal
ANN, Scaled Conjugate Gradient Algorithm		99.1%	Datasets contain SMS spam, DIT spam, British language	feature abstraction, Replacement of similar words, Tokenization, Stemming, Lowercase conversion
Hierarchy linguistic model, Linguistic decision trees		Improve the performance	UCI SMS database	Positive feature, Special Characters removal, Tokenization
NB, Logistic Regression, CNN, LSTM, Random Forest algorithm, The boosted Gradient	CNN	99.44%	Dataset downloaded from the UCI	Feature extraction
The gradient boosted trees, The random forest, NB, The fast large margin, The decision trees, Support vector machine, LSTM, CNN, The hierarchical attention networks the generalized linear model, The gated recurrent unit	RDML	99.26%	Dataset from UCI	Automatically extract by the deep learning classifiers

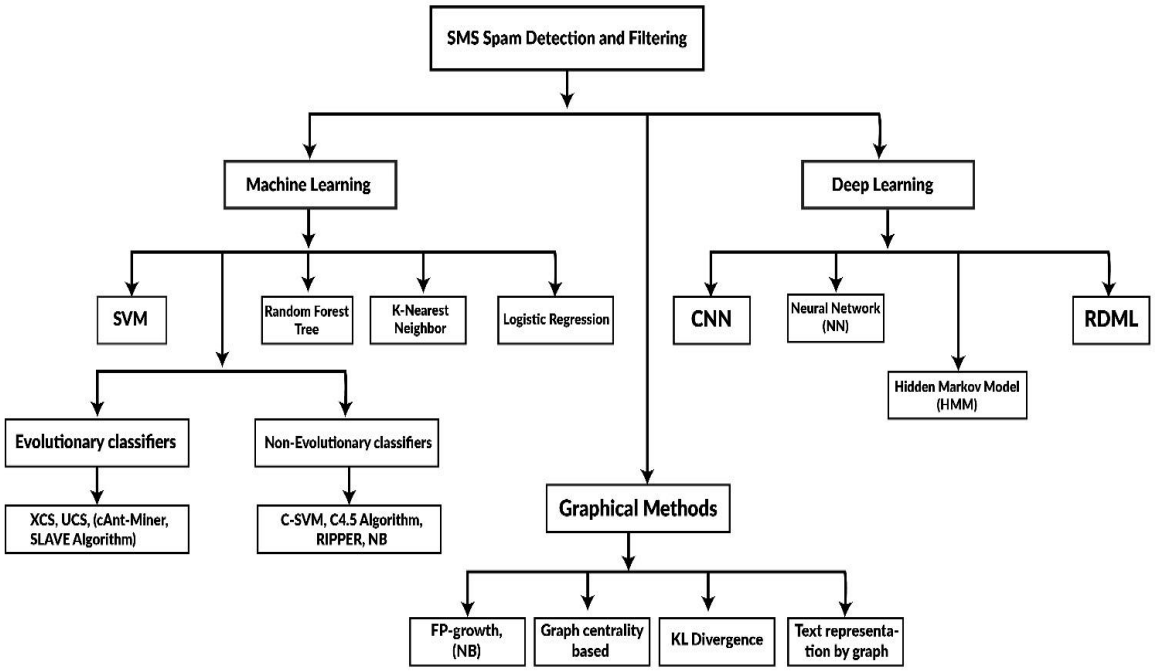

Fig. 8. Spam detection literature model

A method for discretely identifying spam messages on the cell phone survey layer in real world datasets. They make use of two octet-based components: octet bigrams and recursion dispersion of bytes. The directed classifier displays higher exactness of 93% with a no bogus rate alert in a comparison between evolutionary classifiers (the fluffy Ada help, the directed classifier, the hereditary classifier, and the lengthy classifier) and non-developmental classifiers (the K closest neighbour, the Naive Bayes Algorithm, C4.5, Jrip, and SVM) [64,65].

4. CONCLUSIONS

Over the past two decades, a sizable research community has become interested in spam identification and filtration. Many studies have been conducted in this field because to its expensive and significant impact in a variety of situations, including customer behavior and bogus reviews. The survey covers different machine learning methods and models that different researchers have suggested for spam detection and filtering. The study divided them into categories including unsupervised learning, supervised and so forth. The study contrasts different methods and gives an overview of the key takeaways for each group. This study comes to the conclusion that supervised machine learning techniques constitute the foundation of the majority of the proposed spam detection approaches. The supervised model training process depends on a large and time-consuming labelled dataset. SVM and Naive Bayes, supervised learning algorithms, outperform other models in spam identification. The report offers in-depth analyses of these algorithms as well as some suggestions for further research in spam filtering and detection.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Faris H, Al-Zoubi AM, Heidari AA, Aljarah I, Mafarja M, Hassonah MA, et al. An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. Inf Fusion. 2019;48:67-83. DOI: 10.1016/j.inffus.2018.08.002
2. Blanzieri E, Bryl A. A survey of learning-based techniques of email spam filtering. Artif Intell Rev. 2008;29(1):63-92. DOI: 10.1007/s10462-009-9109-6
3. Choudhary K, Garrity KF, Reid ACE, DeCost B, Biacchi AJ, Hight Walker AR, et

al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. npj Comp Mater. 2020;6(1):173.
DOI: 10.1038/s41524-020-00440-1

4. Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. npj Comp Mater. 2015;1(1):1-15.
DOI: 10.1038/npjcompumats.2015.10
5. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Mater. 2013;1(1):011002.
DOI: 10.1063/1.4812323
6. Alghoul A, et al. Email classification using artificial neural network; 2018.
7. Udayakumar N, Anandaselvi S, Subbulakshmi T. Dynamic malware analysis using machine learning algorithm. In: International Conference on Intelligent Sustainable Systems (ICISS). Vol. 2017. IEEE Publications; 2017.
DOI: 10.1109/ISS1.2017.8389286
8. Olatunji SO. Extreme learning machines and support vector machines models for email spam detection. In: 30th Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE Publications. IEEE Publications; 2017.
DOI: 10.1109/CCECE.2017.7946806
9. Dou Y, Ma G, Yu PS, Xie S. Robust spammer detection by nash reinforcement learning. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining; 2020:924-33.
DOI: 10.1145/3394486.3403135
10. Lai G-H, Chen C, Laih C, Chen T. A collaborative anti-spam system. Expert Syst Appl. 2009;36(3):6645-53.
DOI: 10.1016/j.eswa.2008.08.075
11. Dean J. Large scale deep learning. In: Keynote GPU Technical Conference; 2015.
12. Chiu Y-F, Chen C, Jeng B, Lin H. An alliance-based anti-spam approach. In: Third International Conference on Natural Computation (ICNC 2007). IEEE Publications; 2007.
DOI: 10.1109/ICNC.2007.173
13. Smadi S, Aslam N, Zhang L. Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. Decis Support Syst. 2018;107:88-102.
DOI: 10.1016/j.dss.2018.01.001
14. Narisawa K, et al. Unsupervised spam detection based on string alienness measures. in Discovery Science: 10th International Conference. Proceedings, DS 2007 Sendai, Japan. Springer. 2007;10.
15. Sasaki M, Shinnou H. Spam detection using text clustering. In: International Conference on Cyberworlds (CW'05). Vol. 2005. IEEE Publications; 2005.
DOI: 10.1109/CW.2005.83
16. Kruschke JK, Liddell TM. Bayesian data analysis for newcomers. Psychon Bull Rev. 2018;25(1):155-77.
DOI: 10.3758/s13423-017-1272-1, PMID 28405907.
17. Adewole KS, Anuar NB, Kamsin A, Varathan KD, Razak SA. Malicious accounts: Dark of the social networks. J Netw Comput Appl. 2017;79:41-67.
DOI: 10.1016/j.jnca.2016.11.030
18. Zhuang L, et al. Characterizing botnets from email spam records. Leet. 2008; 8(1):1-9.
19. Barushka A, Hájek P. Spam filtering using regularized neural networks with rectified linear units. In: AI* IA 2016 advances in artificial intelligence. XVth International Conference of the Italian Association for Artificial Intelligence, Genova, Italy, November 29 - December 1, 2016, proceedings. Springer; 2016;XV: 65-75.
DOI: 10.1007/978-3-319-49130-1_6
20. Jamil F, Kahng HK, Kim S, Kim DH. Towards secure fitness framework based on IoT-enabled blockchain network integrated with machine learning algorithms. Sensors (Basel). 2021; 21(5):1640.
DOI: 10.3390/s21051640, PMID 33652773.
21. Arif MH, Li J, Iqbal M, Liu K. Sentiment analysis and spam detection in short informal text using learning classifier systems. Soft Comput. 2018;22(21): 7281-91.
DOI: 10.1007/s00500-017-2729-x
22. Zheng X, Zhang X, Yu Y, Kechadi T, Rong C. ELM-based spammer detection in social networks. J Supercomput. 2016;72(8): 2991-3005.
DOI: 10.1007/s11227-015-1437-5
23. Cresci S, Petrocchi M, Spognardi A, Tognazzi S. On the capability of evolved

spambots to evade detection via genetic engineering. Online Soc Netw Media. 2019;9:1-16.
DOI: 10.1016/j.osnem.2018.10.005

24. Saleh AJ, Karim A, Shanmugam B, Azam S, Kannoorpatti K, Jonkman M, et al. An intelligent spam detection model based on artificial immune system. Information. 2019;10(6):209.
DOI: 10.3390/info10060209.
25. Vyas T, Prajapati P, Gadhwal S. A survey and evaluation of supervised machine learning techniques for spam e-mail filtering. In: IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). 2015;2015.
DOI: 10.1109/ICECCT.2015.7226077
26. Jain AK, Gupta BB. Towards detection of phishing websites on client-side using machine learning based approach. Telecommun Syst. 2018;68(4):687-700.
DOI: 10.1007/s11235-017-0414-0
27. Pathan M, Kamble V. A review various techniques for content based spam filtering. Eng Technol. 2018;4.
28. Jain AK, Gupta BB. A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP J Inf Sec. 2016;2016:1-11.
29. Bhowmick A, Hazarika SM. Machine learning for e-mail spam filtering [review]. Techniques and trends. arXiv preprint arXiv:1606.01042, 2016.
30. Bassiouni M, Ali M, El-Dahshan EA. Ham and spam e-mails classification using machine learning techniques. J Appl Sec Res. 2018;13(3):315-31.
DOI: 10.1080/19361610.2018.1463136
31. Ara J. A survey of existing e-mail spam filtering methods considering machine learning techniques. Glob J Comput Sci Technol. 2018;18(C2):21-9.
32. Méndez JR, Cotos-Yañez TR, Ruano-Ordás D. A new semantic-based feature selection method for spam filtering. Appl Soft Comput. 2019;76:89-104.
DOI: 10.1016/j.asoc.2018.12.008
33. Petersen LN. The ageing body in Monty Python Live (Mostly). Eur J Cult Stud. 2018;21(3):382-94.
DOI: 10.1177/1367549417708435
34. Gansterer WN, Janecek AGK, Neumayer R. Spam filtering based on latent semantic indexing. In: Berry MW, Castellanos M, editors. Survey of text mining II: Clustering, classification, and retrieval. London: Springer. London. 2008;165-83.
35. Lee D, Lee MJ, Kim BJ. Deviation-based spam-filtering method via stochastic approach. Europhys Lett. 2018;121(6): 68004.
DOI: 10.1209/0295-5075/121/68004
36. Jain AK, Gupta BB. Towards detection of phishing websites on client-side using machine learning based approach. Telecommun Syst. 2018;68(4):687-700.
DOI: 10.1007/s11235-017-0414-0
37. Ahmed N, Amin R, Aldabbas H, Koundal D, Alouffi B, Shah T. Machine learning techniques for spam detection in Email and IoT platforms: analysis and research challenges. Sec Commun Netw. 2022; 2022:1-19.
DOI: 10.1155/2022/1862888
38. Cabrera-León Y, García Báez P, Suárez-Araujo CP. Non-email spam and machine learning-based anti-spam filters: Trends and some remarks. In: Computer Aided Syst Theor–EUROCAST: 16th International Conference, Las Palmas de Gran Canaria, Spain, Feb 19-24, 2017. Revised selected papers. part I16. Springer. 2017;2018.
39. Subasi A, Alzahrani S, Aljuhani A, Aljedani M. Comparison of decision tree algorithms for spam E-mail filtering. In: 1st International Conference on Computer Applications & Information Security (ICCAIS). 2018;2018.
DOI: 10.1109/CAIS.2018.8442016
40. Hijawi W, Faris H, Alqatawna J, Al-Zoubi AM, Aljarah I. Improving email spam detection using content based feature engineering approach. In: IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). 2017;2017.
DOI: 10.1109/AEECT.2017.8257764
41. DeBarr D, Wechsler H. Using social network analysis for spam detection. In: Advances in social computing. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010;62-9.
DOI: 10.1007/978-3-642-12079-4_10
42. Faris H, Aljarah I, Al-Shboul B. A hybrid approach based on particle swarm optimization and random forests for E-mail spam filtering. In: Computational collective intelligence. Cham: Springer International Publishing. 2016;498-508.
DOI: 10.1007/978-3-319-45243-2_46

43. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerg Artif Intell Appl Comput Eng. 2007;160(1):3-24.
44. Jiang S, Pang G, Wu M, Kuang L. An improved K-nearest-neighbor algorithm for text categorization. Expert Syst Appl. 2012;39(1):1503-9.
DOI: 10.1016/j.eswa.2011.08.040
45. Fine S, Singer Y, Tishby N. The hierarchical hidden markov model: Analysis and applications. Mach Learn. 1998;32(1):41-62.
DOI: 10.1023/A:1007469218079
46. Abe N, Warmuth MK. On the computational complexity of approximating distributions by probabilistic automata. Mach Learn. 1992;9(2-3):205-60.
DOI: 10.1007/BF00992677
47. Baldi P, Chauvin Y, Hunkapiller T, McClure MA. Hidden Markov models of biological primary sequence information. Proc Natl Acad Sci U S A (USA). 1994;91(3):1059-63.
DOI: 10.1073/pnas.91.3.1059, PMID 8302831.
48. Bengio Y, Frasconi P. An input-output HMM architecture. In: Tesauro G, Touretzky DS, Leen TK, editors. Advances in neural information processing systems. Cambridge, MA: MIT Press; 1995.
49. Gat I, Tishby N, Abeles M. Hidden Markov modeling of simultaneously recorded cells in the associative cortex of behaving monkeys. Netw Comput Neural Syst. 1997;8.
50. Cover T, Thomas J. Elements of information theory. Wiley; 1991.
51. Ahmed AH, Mikki M. Improved spam detection using DBSCAN and advanced digest algorithm. Int J Comput Appl. 2013;69(25):11-6.
DOI: 10.5120/12126-8300
52. Tan E, Guo L, Chen S, Zhang X, Zhao Y. Unik: unsupervised social network spam detection. In: Proceedings of the 22nd ACM international conference on information & knowledge management; 2013:479-88.
DOI: 10.1145/2505515.2505581
53. Sharma A, Rastogi V. Spam filtering using K mean clustering with local feature selection classifier. Int J Comput Appl. 2014;108(10):35-9.
DOI: 10.5120/18951-0096
54. Hsiao W-F, Chang T-M. An incremental cluster-based approach to spam filtering. Expert Syst Appl. 2008;34(3):1599-608.
DOI: 10.1016/j.eswa.2007.01.018
55. Ahuja R, Chug A, Gupta S, Ahuja P, Kohli S. Classification and clustering algorithms of machine learning with their applications. In: Nature-inspired computation in data mining and machine learning; 2020. p. 225-48.
DOI: 10.1007/978-3-030-28553-1_11
56. Li W, Meng W, Tan Z, Xiang Y. Design of multi-view based email classification for IoT systems via semi-supervised learning. J Netw Comput Appl. 2019;128:56-63.
DOI: 10.1016/j.jnca.2018.12.002
57. Diale M, Celik T, Van Der Walt C. Unsupervised feature learning for spam email filtering. Comput Electr Eng. 2019;74:89-104.
DOI: 10.1016/j.compeleceng.2019.01.004
58. Peng W, Huang L, Jia J, Ingram E. Enhancing the naive Bayes spam filter through intelligent text modification detection. In: 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications. 2018;2018.
DOI:10.1109/TrustCom/BigDataSE.2018.00122
59. Zeng Z, et al. Spammer Detection on Weibo Social Network. In: IEEE 6th International Conference on Cloud Computing Technology and Science. 2014;2014.
60. Lei K, Liu Y, Zhong S, Liu Y, Xu K, Shen Y, et al. Understanding user behavior in Sina Weibo online social network: A community approach. IEEE Access. 2018;6:13302-16.
DOI: 10.1109/ACCESS.2018.2808158
61. Lin C, He J, Zhou Y, Yang X, Chen K, Song L. Analysis and identification of spamming behaviors in Sina Weibo microblog. In: Proceedings of the 7th workshop on social network mining and analysis. Chicago: Association for Computing Machinery. 2013:Article 5.
DOI: 10.1145/2501025.2501035
62. Rusland, N.F., et al. Analysis of naïve Bayes algorithm for Email spam filtering across multiple datasets. IOP Conf S Mater Sci Eng. 2017;226(1):012091.
63. Singh A, Batra S. Ensemble based spam detection in social IoT using probabilistic data structures. Future Gener Comput Syst. 2018;81:359-71.
DOI: 10.1016/j.future.2017.09.072

64. Xu H, Sun W, Javaid A. Efficient spam detection across Online Social Networks. In: IEEE International Conference on Big Data Analysis (ICBDA). 2016;2016.
DOI: 10.1109/ICBDA.2016.7509829
65. Faris H, Aljarah I, Alqatawna J. Optimizing feedforward neural networks using Krill Herd algorithm for E-mail spam detection. In: IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT). 2015;2015.
DOI: 10.1109/AEECT.2015.7360576
66. Opera: state of the mobile [web]. Available:http://www.opera.com/smw/2009/12
67. Sahami M, Dumais S, Heckerman D, Horvitz E. A bayesian approach to filtering junk e-mail. In: AAAI Workshop on Learning for Text Categorization; 1998.
68. Gyöngyi Z, Garcia-Molina H, Pedersen J. Combating web spam with trustrank. In: Proceedings of the thirtieth international conference on very large data bases. 2004;576-87.
69. Gyongyi Z, Berkhin P, Garcia-Molina H, Pedersen J. Link spam detection based on mass estimation. In: VLDB 2006. Proceedings of the 32nd international conference on very large data bases. 2006;439-50.
70. Zhou D, Burges CJC, Tao T. Transductive link spam detection. In: Proceedings of the 3rd international workshop on adversarial information retrieval on the web. 2007;21-8.
DOI: 10.1145/1244408.1244413
71. Geng GG, Li Q, Zhang X. Link based small sample learning for web spam detection. In: Proceedings of the 18th international conference on world wide web. 2009; 1185-6.
DOI: 10.1145/1526709.1526920
72. Wu Y-S, Bagchi S, Singh N, Wita R. Spam detection in voice-over-ip calls through semi-supervised clustering. In: Proceedings of the. Dependable systems networks. 2009;307-16.
DOI: 10.1109/DSN.2009.5270323
73. Benevenuto F, Rodrigues T, Almeida V, Almeida J, Gonçalves M. Detecting spammers and content promoters in online video social networks. In: Proceedings of the 32nd international ACM SIGIR conference. 2009;620-7.
DOI: 10.1145/1571941.1572047
74. Krishnamurthy B, Gill P, Arlitt M. A few chirps about twitter. In: Proceedings of the first workshop on online social networks. 2008;19-24.
DOI: 10.1145/1397735.1397741