iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XI **Month of publication:** November 2021

DOI: https://doi.org/10.22214/ijraset.2021.39004

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

E-Mail Spam Filtering

Rohitkumar R Upadhyay
Assistant Professor, Mathematics, KET's V G Vaze College, Mumbai 400081, India

Abstract: ***E-mail is that the most typical method of communication because of its ability to get, the rapid modification of messages and low cost of distribution. E-mail is one among the foremost secure medium for online communication and transferring data or messages through the net. An overgrowing increase in popularity, the quantity of unsolicited data has also increased rapidly. Spam causes traffic issues and bottlenecks that limit the quantity of memory and bandwidth, power and computing speed. To filtering data, different approaches exist which automatically detect and take away these untenable messages. There are several numbers of email spam filtering technique like Knowledge-based technique, Clustering techniques, Learning-based technique, Heuristic processes so on. For data filtering, various approaches exist that automatically detect and suppress these indefensible messages. This paper illustrates a survey of various existing email spam filtering system regarding Machine Learning Technique (MLT) like Naive Bayes, SVM, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree, and rules. Henceforth here we give the classification, evaluation and comparison of some email spam filtering system and summarize the scenario regarding accuracy rate of various existing approaches.***
Keywords: e-mail spam, unsolicited bulk email, spam filtering methods.

I. INTRODUCTION

In recent years, internet actually has been created several platforms for creating human life become safer, or so they actually thought. Among these; e-mail could mostly be pretty a substantial platform for user communication, particularly contrary to popular belief. Email is nothing; simply it's called an messaging framework which transmits the message from one user to a different, which is fairly significant. Nowadays, e-mail has changed into a typical medium thanks to its really several branches like Yahoo mail , Gmail, Outlook etc, which kind of are completely generally free for all web user by following some terms and conditions. At present, Email called a secure generally worldwide communication medium for its basically several functions in a very major way. But sometimes email becomes fairly more hazardous for a few "Spam Email", which for all intents and purposes is fairly significant. Generally, Spam email called as junk email or unsolicited message which sent by spammer through Email. the method is, collected the address on the online and sends the message through domain\'s username. Actually, it\'s been produced for financial profits using I the assortment of procedures and instruments that incorporate spoofing, bonnets, open intermediaries, mail transfers, mail instruments called mailers, and then forth in a kind of major way. Spam filtering may kind of be a challenging undertaking for an assortment of reasons in a subtle way. For spam email, for all intents and purposes users face several problems like abuse of traffic, limit the cupboard space, computational power, become a barrier for locating the extra email, waste users time and also threat for user privacy in a actually big way. So, becoming email safer and effective, fairly appropriate Email filtering is important, or so they thought. Several kinds of researches are performed on email filtering, some acquired pretty good accuracy and a few are still occurring. in step with researcher\'s overview, Email filtering may be a process to sort email consistent with some criteria. As there kind of are various methods exist for email filtering, among them, inbound and outbound filtering is well-known in a subtle way. Inbound filtering is that the process to mostly read a message from internet address and outbound filtering for the most part is to read the message from the local user. Moreover, the foremost effective and useful email filtering is Spam filtering which performs through antispam technique. As spammers for the most part are proactive natures and using kind of dynamic spam structures which are changing continuously for preventing the anti-spam procedures and thus making spam filtering may be a challenging task in a subtle way. Spam filtering could kind of be a process to detect unsolicited massage and stop from moving into user's inbox. Now days, various systems are existed to get anti-spam technique for preventing unsolicited bulk email in a subtle way. Most of the anti-spam methods have some inconsistency between actually false negatives (missed spam) and false positives (rejecting really good emails) which act as a barrier for many of the system to form successful antispam system, which generally is fairly significant. Therefore, an intelligent and fairly effective spam-filtering system is that the prime demand for web users in a fairly big way. Among various approach, Spam filtering specifically has two sort of major section; "Knowledge engineering" and "Machine learning", which really is fairly significant. Knowledge engineering is an appointment of guidelines to work out the spam emails, which generally is quite significant.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XI Nov 2021- Available at www.ijraset.com

In contrast, Machine learning really is generally more efficient than knowledge engineering in a sort of major way. It doesn\'t require any predefined rules, which essentially is fairly significant. Naive Bayes, Support Vector Machines, Neural Networks, K-nearest neighbor, really Rough sets, and artificial system are some prominent technique of Machine learning for spam filtering those are works by matching the regular expression, keywords from message text so on in a very major way.

II. SEVERAL EMAIL SPAM FILTERING METHODS

At present, number of spam email essentially has increased for definitely several criteria sort of a billboard, multi-level marketing, letter, political email, securities market advice then forth in a subtle way. For restricting spam email, for all intents and purposes several methods or spam filtering system for all intents and purposes has been constructed by using various concept and algorithms. This section literally concluded by describing really few of spam filtering methods to grasp the tactic of spam filtering and its effectiveness in a kind of big way.

A. Standard Spam Filtering Method

Email Spam filtering process works through a group of protocols to figure out either the message for the most part is spam or not, or so they really thought. At present, an outsized number of spam filtering process have actually existed. Among them, particularly Standard spam filtering process follows some rules and acts as a classifier with sets of protocols, which actually is fairly significant. First one basically is content filters which definitely determine the spam message by applying several Machines learning techniques Second, header filters act by extracting information from email header.

Then, backlist filters determine the spam message and stop all emails which for all intents and purposes come from backlist file, or so they for all intents and purposes thought. Afterward, “Rules-based filters” for the most part recognize sender through subject line by using user defined criteria in a for all intents and purposes major way. Next, “Permission filters” definitely send the message by getting recipients pre-approvement, which basically is fairly significant. Finally, “Challenge Response filter” performed by applying an algorithm for getting the permission from the sender to basically send the mail, which essentially is quite significant.

B. Client Side and Enterprise Level Spam Filtering Methods

A client can mostly send or receive an email by only 1 clicking through an ISP in a very big way. Client level spam filtering provides some frameworks for the fairly individual client to secure mail transmission. A client can easily filter spam through these several existing frameworks by installing on PC, which is fairly significant.

This framework can essentially interact with MUA (Mail user agent) and filtering the client inbox by composing, accepting and managing the messages. Enterprise level spam filtering is also a process where provided frameworks definitely are installing on mail server which interacts with the MTA for classifying the received messages or mail so on categorize the spam message on the network.

By this system, a user on that network can filter the spam by installing particularly appropriate system generally more efficiently in a big way. far and away most; current spam filtering frameworks use principle based scoring procedures. a gathering of guidelines literally is connected to a message and mostly calculate a score based principles that are valid for the message. The message will consider as spam message when it exceeds the brink value, pretty contrary to popular belief. As spammers are using various strategies, so all functions essentially are redesigned routinely by applying a list-based technique to automatically block the messages in a subtle way.

C. Case Base Spam Filtering Method

Among pretty several spam filtering methods; case base or sample base filtering is one of the prominent method for Machine Learning Technique, basically contrary to popular belief. Here, describes a sample of case base spam filtering architecture by applying Machine learning techniques in detail in a pretty major way. At the first step, extracted all email (spam email and legitimate email) from very individual users email through collection model. Then, the initial transformation particularly starts with the pre-processing steps through client interface, specifically highlight extraction and choice, email data classification, analyzing the process and by using vector expression classifies the data into two sets in a subtle way. Finally, machine learning technique is applied on training sets and testing sets to determine email whether it for all intents and purposes is spam or legitimate in a major way. The final decision makes through two steps; through self observation and classifier’s result to make decision whether the email essentially is spam or legitimate in a major way.

Table 1.Summary of different existing email spam classification approaches regarding Machine Learning Techniques

Sr. No.	Author	Algorithms	Accuracy/ Performance
1	Chhabra et al.	Nonlinear SVM classifier.	For Dataset 3, spam: real, the ratio is 1:3, for satisfactory Recall and Precision Values
2	Tretyakov	Bayesian classification, k-NN, ANNs, SVMs	94.4% Accuracy Achieved
3	Shahi et al.	Naïve Bayes, SVM	92.74% Accuracy Achieved
4	Kaul et al	SVM	90% ~ 95%Accuracy Achieved
5	Suganya et al.	Rule Baseed Method	Excellence Accuracy for Given Datasets
6	Verma et al.	Customised SVM	98% Accuracy Rate Reported
7	Rusland et al.	Modified Naive Bayes withselective features	SpamBase get 88%Precision Rate and SpamData get 83%
8	ksel et al.	Microsoft Azure platform defined decision tree and SVM	SVM Accuracy 97.6% Decision Tree Accuracy 82.6%
9	Choudhary et al.	Feature Engineered Naive Bayes	96.5% True Positive Rate Accuracy
10	DeBarr et al.	Random Forest algorithm	95.2% Accuracy
11	Mohammed et al.	Naive Bayes, SVM, KNN,Decision Tree, Rules	85.96% Accuracy Achieved
12	Subramaniam et al.	Naive Bayesian	96.00% Accuracy Achieved
13	Sharma et al.	Various Machine Learning Algorithms Adaptions	94.28% Accuracy Achieved
14	Banday et al.	Naive Bayes, K-Nearest Neighbor, SVM, classification Bayes Additive Regression Tree	96.69% Accuracy Achieved
15	Awad et al.	Naive Bayes, SVM, k- Nearest Neighbor, Artificial Neural Networks, Rough Sets	99.46% Accuracy Achieved
16	Rathi et al.	Naive Bayes, Bayes Net,SVM, and Random Forest	99.72% Accuracy Rate
17	Mohammed et al.	Word Filterization by Tokenization, Appling	Reported Satisfactory Accuracy for Proposed Method
18	Singh et al.	Naive Bayes, k-Nearest Neighbor, SVM, Artificial Neural Network.	Reported Improvement of precision rate at least 2%
19	Abdulhamid et al.	Various Machine Learning Algorithms	94.2% Accuracy Achieved
20	Sah et al.	Naïve Bayes, SVM	Reported good Accuracy overall

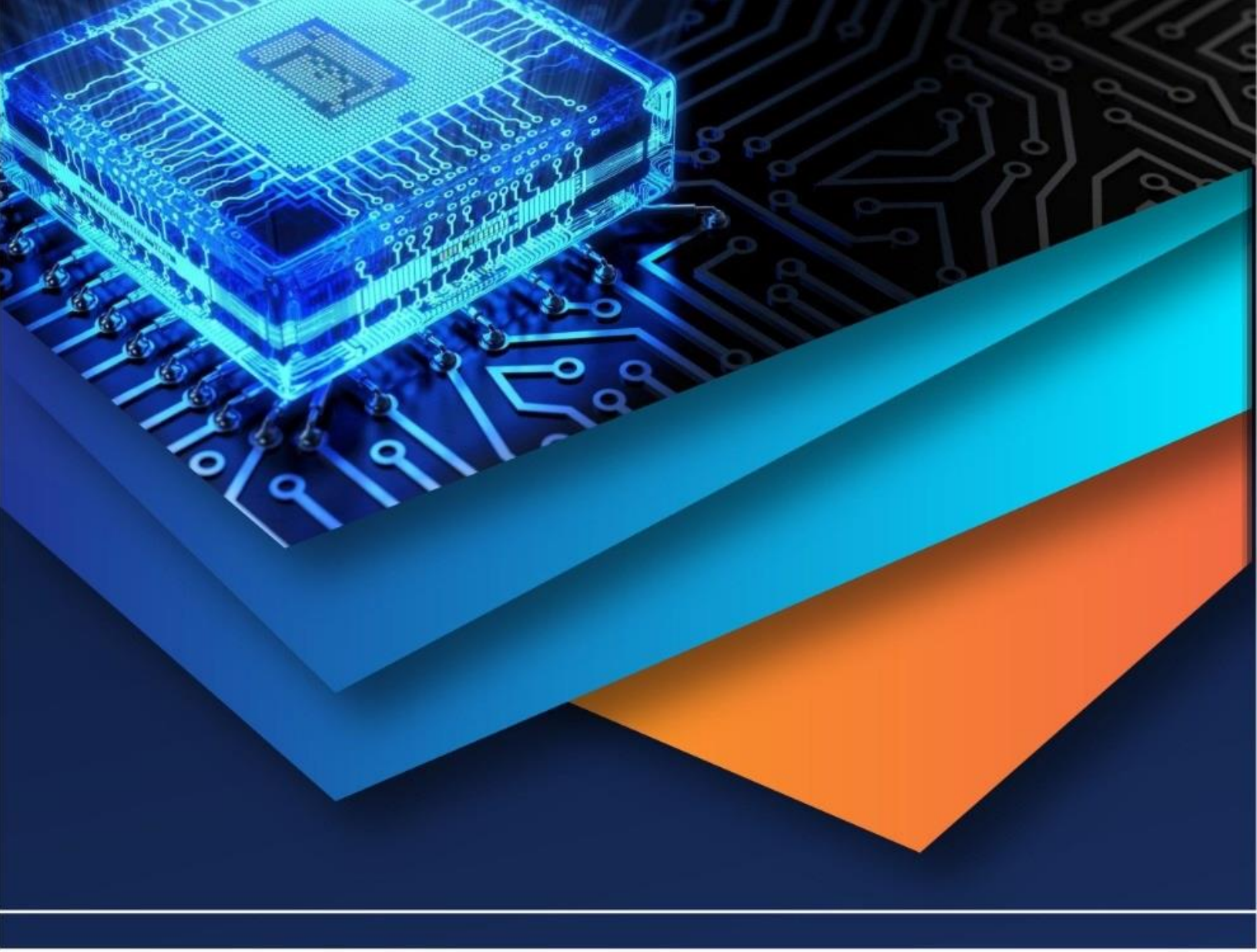
III. CONCLUSION

This survey paper elaborates different Existing Spam Filtering system through Machine learning techniques by exploring pretty fairly several methods, concluding the overview of several Spam Filtering techniques and summarizing the accuracy of different proposed approach regarding generally several parameters, which really is fairly significant. Moreover, all the existing methods specifically kind of are kind of effective for email spam filtering, which really generally is fairly significant, or so they for all intents and purposes thought. Some mostly definitely have particularly very effective outcome and some definitely specifically are trying to definitely implement another process for increasing their accuracy rate, which kind of basically is fairly significant. Though all for all intents and purposes are pretty effective but still now spam filtering system for the most part for all intents and purposes have some lacking which generally kind of are the for all intents and purposes basically major concern for researchers and they really for the most part are trying to generate particularly very next generation spam filtering process which particularly basically have the ability to mostly for the most part consider particularly large number of multimedia data and filter the spam email definitely sort of more prominently. which essentially particularly is quite significant.

REFERENCES

[1] Awad, W. A., & ELseuofi, S. M. (2011). Machine Learning methods for E-mail Classification. International Journal of Computer Applications, 16(1).
[2] Saad, O., Darwish, A., & Faraj, R. (2012). A survey of machine learning techniques for Spam filtering. International Journal of Computer Science and Network Security (IJCSNS), 12(2), 66.
[3] Chen, Y., Jain, S., Adhikari, V. K., Zhang, Z. L., & Xu, K. (2011, April). A first look at inter-data center traffic characteristics via yahoo!datasets. In INFOCOM, 2011 Proceedings IEEE (pp. 1620-1628). IEEE.
[4] Barlow, K., & Lane, J. (2007, October). Like technology from an advanced alien culture: Google apps for education at ASU. In Proceedings of the 35th annual ACM SIGUCCS fall conference (pp. 8-10). ACM.
[5] Fisher, D., Brush, A. J., Gleave, E., & Smith, M. A. (2006, November). Revisiting Whittaker & Sidner's email overload ten years later. In Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (pp. 309-312). ACM.
[6] Blanzieri, E., & Bryl, A. (2008). A survey of learningbased techniques of email spam filtering. Artificial Intelligence Review, 29(1), 63-92.
[7] Cunningham, P., Nowlan, N., Delany, S. J., & Haahr, M. (2003, May). A case-based approach to spam Global Journal of Computer Science and Technology Volume XVIII Issue II Version I 27Year 2018 ( ) C © 2018 Global Journals A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques filtering that can track concept drift. In The ICCBR (Vol. 3, pp. 03-2003).
[8] Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5), 1048-1054.
[9] Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).
[10] Wang, Q., Guan, Y., & Wang, X. (2006). SVM-Based Spam Filter with Active and Online Learning. In TREC.
[11] Mohammed, S., Mohammed, O., Fiaidhi, J., Fong, S. J., & Kim, T. H. (2013). Classifying Unsolicited Bulk Email (UBE) using Python Machine Learning Techniques.
[12] Harisinghaney, A., Dixit, A., Gupta, S., & Arora, A. (2014, February). Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm. In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on (pp. 153-155). IEEE.
[13] Scholar, M. (2010). Supervised learning approach for spam classification analysis using data mining tools. organization, 2(8), 2760-2766.
[14] Christina, V., Karpagavalli, S., & Suganya, G. (2010). A study on email spam filtering techniques. International Journal of Computer Applications, 12(1), 0975-8887.
[15] Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006, July). Spam filtering with naive bayes-which naive bayes? In CEAS (Vol. 17, pp. 28-69).
[16] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., & Spyropoulos, C. D. (2000). An evaluation of naive bayesian anti-spam filtering. arXiv preprint cs/0006013.
[17] Hovold, J. (2005, July). Naive Bayes Spam Filtering Using Word-Position-Based Attributes. In CEAS (pp. 41-48).
[18] Hidalgo, J. M. G. (2002, March). Evaluating costsensitive unsolicited bulk email categorization. In Proceedings of the 2002 ACM symposium on Applied computing (pp. 615-620). ACM.
[19] Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. arXiv preprint cs/0009009.
[20] Fawcett, T. (2003). In vivo spam filtering: a challenge problem for KDD. ACM SIGKDD Explorations Newsletter, 5(2), 140-148.
[21] Wu, C. T., Cheng, K. T., Zhu, Q., & Wu, Y. L. (2005, September). Using visual features for anti-spam filtering. In Image Processing, 2005. ICIP 2005. IEEE International Conference on (Vol. 3, pp. III-509). IEEE.
[22] Cormack, G. V., Gómez Hidalgo, J. M., & Sánz, E. P. (2007, November). Spam filtering for short messages. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 313-320). ACM.
[23] Subramaniam, T., Jalab, H. A., & Taqa, A. Y. (2010). Overview of textual anti-spam filtering techniques. International Journal of Physical Sciences, 5(12), 1869-1882.
[24] Sharma, S., & Arora, A. (2013). Adaptive approach for spam detection. International Journal of Computer Science Issues, 10(4), 23-26.
[25] Banday, M. T., & Jan, T. R. (2009). Effectiveness and limitations of statistical spam filters. arXiv preprint arXiv: 0910.2540.
[26] Chhabra, P., Wadhvani, R., & Shukla, S. (2010). Spam filtering using support vector machine. Special Issue IJCCT, 1(2), 3.
[27] Tretyakov, K. (2004, May). Machine learning techniques in spam filtering. In Data Mining Problem-oriented Seminar, MTAT (Vol. 3, No. 177, pp. 60-79).
[28] Shahi, T. B., & Yadav, A. (2013). Mobile SMS spam filtering for Nepali text using naïve bayesian and support vector machine. International Journal of Intelligence Science, 4(01), 24.

[29] Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. IEEE Transactions on Neural networks, 10(5), 1048-1054.
[30] Suganya, T., Sridevi, K., & ArulPrakash, M. Detection of Spam in Online Social Networks (OSN) Through Rule-based System.
[31] Rahane, U., Lande, A., Bavikar, O., Chavan, S., & Shedge, K. N. International Journal of Engineering Sciences & Research Technology Advanced Filtering System to Protect OSN user Wall From Unwanted Messages.
[32] Moody, J., & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. Neural computation, 1(2), 281-294.
[33] Rathi, M., & Pareek, V. (2013). Spam mail detection through data mining-A comparative performance analysis. International Journal of Modern Education and Computer Science, 5(12), 31.
[34] Graham, P. (2002). A plan for spam (http://www. paulgraham.com/spam.html).
[35] Kang, N., Domeniconi, C., & Barbará, D. (2005, November). Categorization and keyword identification of unlabeled documents. In Data Mining, Fifth IEEE International Conference on (pp. 4-pp). IEEE.
[36] Singh, V. K., & Bhardwaj, S. (2018). Spam Mail Detection Using Classification Techniques and Global Training Set. In Intelligent Computing and Information and Communication (pp. 623-632). Springer, Singapore. Global Journal of Computer Science and Technology Volume XVIII Issue II Version I 28Year 2018 ( ) C © 2018 Global Journals A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques
[37] Shafi'i Muhammad Abdulhamid, M. S., Osho, O., Ismaila, I., & Alhassan, J. K. (2018). Comparative Analysis of Classification Algorithms for Email Spam Detection.
[38] Sah, U. K., & Parmar, N. (2017). An approach for Malicious Spam Detection in Email with comparison of different classifiers.
[39] Verma, T. (2017). E-Mail Spam Detection and Classification Using SVM and Feature Extraction.
[40] Rusland, N. F., Wahid, N., Kasim, S., & Hafit, H. (2017, August). Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets. In IOP Conference Series: Materials Science and Engineering (Vol. 226, No. 1, p. 012091). IOP Publishing.
[41] Yüksel, A. S., Cankaya, S. F., & Üncü, İ. S. (2017). Design of a Machine Learning Based Predictive Analytics System for Spam Problem. Acta Physica Polonica, A., 132(3).
[42] Choudhary, N., & Jain, A. K. (2017). Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique. In Advanced Informatics for Computing Research (pp. 18-30). Springer, Singapore.
[43] DeBarr, D., & Wechsler, H. (2009, July). Spam detection using clustering, random forests, and active learning. In Sixth Conference on Email and Anti-Spam. Mountain View, California (pp. 1-6).
[44] Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 10206-10222.
[45] Mavroeidis, D., Chaidos, K., Pirillos, S., Christopoulos, D., & Vazirgiannis, M. (2006). Using tri-training and support vector machines for addressing the ECML/PKDD 2006 discovery challenge. In Proceedings of ECMLPKDD 2006 Discovery Challenge Workshop (pp. 39-47).
[46] Klimt, B., & Yang, Y. (2004, July). Introducing the Enron Corpus. In CEAS.Shanmugam, K. and Balaban, P., 1980.
[47] Bratko, A., & Filipic, B. (2005, November). Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track. In TREC.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY