**Email Based Spam Detection**

Dept. of Computer Science and Application,

Gour Mahavidyalaya, Mangalbari, Malda,

West Bengal, India

**Abstract:->>**

E-mail is one of the most secure medium for online communication and transferring data or messages through the web. An overgrowing increase in popularity, the number of unsolicited data has also increased rapidly. To filtering data, different approaches exist which automatically detect and remove these untenable messages. There are several numbers of email spam filtering technique such as Knowledge-based technique, Clustering techniques, Learning-based technique, Heuristic processes and so on. This paper illustrates a survey of different existing email spam filtering system regarding Machine Learning Technique (MLT) such as Naive Bayes, K-Nearest Neighbour, Bayes Additive Regression, KNN Tree, and rules. However, here we present the classification, evaluation and comparison of different email spam filtering system and summarize the overall scenario regarding accuracy rate of different existing approaches.

**Keywords:->>**

Blank

**1 Introduction:>>**

Spam is digital junk mail and unsolicited communications sent in bulk through an electronic messaging system. Unrequested, disruptive, and usually promotional, spam messages are designed to flood as many inboxes as possible. Traditionally, spam has been sent via email, but also includes SMS and social media messaging.

Spammers use many forms of communication to bulk-send their unwanted messages. Some of these are marketing messages peddling unsolicited goods. Other types of spam messages can spread malware, trick you into divulging personal information, or scare you into thinking you need to pay to get out of trouble. Email spam filters catch many of these types of messages, and phone carriers often warn you of a “spam risk” from unknown callers. Whether via email, text, phone, or social media, some spam messages do get through, and you want to be able to recognize them and avoid these threats.

Phishing emails are a type of spam cybercriminals send to many people, hoping to “hook” a few people. Phishing emails trick victims into giving up sensitive information like website logins or credit card information.

In a tech support scam, the spam message indicates that you have a technical problem and you should contact tech support by calling the phone number or clicking a link in the message. Like email spoofing, these types of spam often say they are from a large technology company like Microsoft or a cybersecurity company like Malwarebytes.

Hot topics in the news can be used in spam messages to get your attention. In 2020 when the world was facing the Covid-19 pandemic and there was an increase in work-from-home jobs, some scammers sent spam messages promising remote jobs that paid in Bitcoin. During the same year, another popular spam topic was related to offering financial relief for small businesses, but the scammers ultimately asked for bank account details. News headlines can be catchy, but beware of them in regards to potential spam messages.

Beyond the annoyance and the time wasted sifting through unwanted messages, spam can cause significant harm by infecting users’ computers with malicious software capable of damaging systems and stealing personal information. It also can consume network resources.

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

The machine learning technique uses statistical methods for automatic email classification to filter spam from a user’s inbox. Some popular Machine Learning techniques for spam filtering are Naive Bayes, Support Vector Machines, Decision Trees, Neural Networks, etc. The sophistication of Machine learning algorithms makes it one of the best spam filter services among all other spam filtering techniques. The success of spam filter Gmail can be ascribed to Gmail’s timely transition and successful use of the Machine Learning techniques to filter both, incoming spam and other abuses like Denial-of-Service (DoS) attacks. The reason for the success of Machine Learning based spam filters is that they retrain themselves when put in use and minimize the manual effort required while delivering superior filtering accuracy.

***Literature Survey:***

***Methodology:->>***

Email spam classification is a big problem in today's generation computerized environment. A unique spam class method are used to work around this problem. Using spam detection approach, we can quickly recognise spam and ham emails in our mailbox. In this paper we used supervised and unsupervised learning, as a supervised we used Naïve Bayes, SVM, Random Forest, etc methods. and unsupervised KNN,DT,Adaboost,Xgboost. Its management efficiency and its similarity with its great predictive accuracy have made it a target machine for various machine learning researchers in today's epoch of vital statistics. All the methods has been used to solve many complex issues in relevant systems, overall performance, accuracy, precision, and essential generalizability. Its widespread use, spam detection frameworks performance, and detailed study have yet to be examined to satisfy our experience.

In this paper we also provide SVM, KNN, NB, DT, LR, RF, AdaBoost, BgC, ETC, GBDT, Xgboost

***Dataset Description:->>***

In this paper, we are using ‘spam ham dataset’ from kaggle and the dataset includes total 5572 records with 2 columns (Category as a target and Massage as a text). in this dataset there is no null value and after checking duplicate values we got 454 duplicate values. After removing all duplicate and null values now we have Total Dataset 5157 and 2 columns. In the dataset email was considered spam (1) or ham (0). as a target columns we have total ham 4516 and spam 641 (87.57% ham and 12.43% spam). The Dataset divided into two categories 80% is training set and 20% is testing set. The experimental conducted on the needed dataset of both training set and testing set.

***Result:->>***

|  |  |  |  |
| --- | --- | --- | --- |
| **Reference** | **Algorithm** | **Accuracy** | **Proposed Accuracy** |
| PAPER 1 | SVM | 94.06% | 97.00% |
| NB | 97.47% | 97.67% |
| XGBoost | 95.00% | 97.38% |
| PAPER 2 | KNN | 82.34% | 90.79% |
| PAPER 3 | RF | 95.40% | 97.77% |
| ETC | 95.50% | 96.96% |
| DT | 92.10% | 93.50% |
| PAPER 4 | AdaBoost | 95.92% | 97.09% |

***CONCLUSIONS:->>***

This survey paper elaborates different Existing Spam Filtering system through Machine learning techniques by exploring several methods, concluding the overview of several Spam Filtering techniques and summarizing the accuracy of different proposed approach regarding several parameters. Moreover, all the existing methods are effective for email spam filtering. Some have effective outcome and some are trying to implement another process for increasing their accuracy rate. Though all are effective but still now spam filtering system have some lacking which are the major concern for researchers and they are trying to generate next generation spam filtering process which have the ability to consider large number of multimedia data and filter the spam email more prominently and server filtering systems is rather low. But it can be improved if to apply the hybrid filtration system in other words the complex hierarchical and multi-agent filtration system that helps users to participate in the identification of the filtering errors and the appropriate setting of filters at each level (user level, organization level, mail provider level).

Therefore it is quite perspective for solving this problem, the combination of two widespread approaches as using the personal e-mail classification model on a server side solution. Development of server side personalized e-mail filtering systems that use the learning-based classification algorithms based on Data Mining methods is a very perspective direction.

***Acknowledgements :->>***

***References***

[1] Mustapha, Ismail B., et al. "Effective Email Spam Detection System using Extreme Gradient Boosting." *arXiv preprint arXiv:2012.14430* (2020).

[2] Goyal, Saumya, R. K. Chauhan, and Shabnam Parveen. "Spam detection using KNN and decision tree mechanism in social network." *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. IEEE, 2016.

[3] Zavvar, Mohammad, Meysam Rezaei, and Shole Garavand. "Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine." *International Journal of Modern Education and Computer Science* 8.7 (2016): 68.

[4] Metsis, Vangelis, Ion Androutsopoulos, and Georgios Paliouras. "Spam filtering with naive bayes-which naive bayes?." *CEAS*. Vol. 17. 2006.

[5] Akinyelu, Andronicus A., and Aderemi O. Adewumi. "Classification of phishing email using random forest machine learning technique." *Journal of Applied Mathematics* 2014 (2014).

[6] Devi, Khongbantabam Susila, and R. Ravi. "A new feature selection algorithm for Efficient Spam Filtering using Adaboost and Hashing techniques." *Indian Journal of Science and Technology* 8.13 (2015): 1-8.

[7] Rajamohana, S. P., K. Umamaheswari, and S. Vasantha Keerthana. "An effective hybrid cuckoo search with harmony search for review spam detection." *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. IEEE, 2017.