IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Email Classification using Machine Learning

Prof. R. Y. Thombare, Prathamesh Patil, Pranav Shelar, Omkar Dhakane, Krishna Chakor
Department of Computer Technology
K. K. Wagh Polytechnic, Nashik, Maharashtra, India

Abstract: *Email usage has become a fundamental means of communication for both businesses and personal use. Its widespread usability has also led to an increase in the volume of email data. While e-mails are necessary for everyone, they also come with unnecessary, undesirable bulk mails, which are also called Spam Mails. Email sorting has become a problem due to volumes and if not properly done could lead to inefficiency at work . Anyone with access to the internet can receive spam on their devices. Most spam emails divert people's attention away from genuine and important emails and direct them towards detrimental situations. Spam emails are capable of filling up inboxes or storage capacities, deteriorating the speed of the internet to a great extent. These emails have the capability of corrupting one's system by smuggling viruses into it, or steal useful information and scam gullible people. The identification of spam emails is a very tedious task and can get frustrating sometimes. Wasting time searching through unsorted emails could lead to vulnerability of spam and phishing attacks during that process. This project looks at a comparative research of Naïve Bayes, Neural Networks and SVM performance to classify emails.*

Keywords: Machine learning techniques, Neural Networks, Support Vector Machine, Naïve Bayes

I. INTRODUCTION

Email has become one of the most important forms of communication. In 2014, there are estimated to be 4.1 billion email accounts worldwide, and about 196 billion emails are sent each day worldwide. Spam is one of the major threats posed to email users. In 2013, 69.6% of all email flows were spam. Therefore, an effective spam filtering technology is a significant contribution to the sustainability of the cyberspace and to our society. The Internet has become an inseparable part of human, where more than four and half billion Internet users find it convenient to use it for their facilitation. Moreover, emails are considered as a reliable form of communication by the Internet users. Over the decades, e-mail services have been evolved into a powerful tool for the exchange of different kind of information. The increased use of the e-mail also entails more spam attacks for the Internet users. Spam can be sent from anywhere on the planet from users having deceptive intentions that has access to the Internet. Spams are unsolicited and unwanted emails sent to recipients who do not want or need them. These spam emails have fake content with mostly links for phishing attacks and other threats, and these emails are sent in bulk to a large number of recipients. The intention behind them is to steal users' personal information and then use them against their will to gain materialistic benefits. These emails either contain malicious content or have URLs that lead to malicious content. Such emails are also sometimes referred to as phishing emails. Despite the advancement of spam filtering applications and services, there is no definitive way to distinguish between legitimate and malicious emails because of the ever-changing content of such emails. Spams have been sent forever three or four decades now, and with the availability of various anti-spam services, even today, nonexpert end-users get trapped into such hideous pitfall. In e-mail managers, spam filters detect spam and forward it to a dedicated space, spam folder, allowing the user to choose whether or not to access them. Spam filtering tools such as corporate e-mail, e-mail filtering gateways, contracted antispam services, and end-user training can deal with spam emails in English or any other language.

II. PROJECT CONCEPT AND WORKING

The spam detection engine should be able to take email datasets as input and with the help of text mining and optimized supervised algorithms; it should be able to classify the email as ham or spam.

In this Machine Learning Project, the Algorithms to be used for Spam detection are Support Vector Machine, Random Forest, K-Nearest Neighbours and Decision Tree. The dataset used for this project will consist of spam and ham (normal) emails. The dataset is split into 2 sub-datasets; say "train dataset" and "test dataset" in the proportion of 75:25.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Then the "train" and "test" dataset are then passed as parameters for text-processing. The extracted text is converted into vector for further processing. Then the model is trained for the respective algorithms using the train dataset. Once the model is ready, it is tested on the test dataset. The performance of each algorithm is then analysed using data visualization and performance metrics depicting accuracy of classifying spam and ham emails. This approach helps to conclude which algorithm has the highest accuracy.

III. AREA OF PROJECT

Artificial intelligence and Machine learning, Mailing software's, Organizations, Job portals

The Algorithms used for the project are:

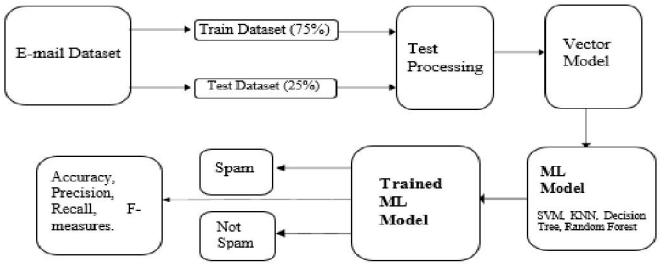
- **Support Vector Machine:** SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane.
- **Random Forest:** Random Forest algorithm is a supervised learning algorithm which is used for both regression and classification. A Random Forest is a meta estimator that fits a number of Decision Tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree. It is an improved version of the Decision Tree
- **K-Nearest Neighbour:** K-Nearest Neighbours algorithm: K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique and it can be used for Regression as well as for Classification. But mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- **Decision Tree:** Decision tree is a supervised learning algorithm. It is used for both regression and classification. The goal of using a decision tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules developed from training data.

IV. DESIGN CONCEPT

In this E-mail classification Algorithms to be used for Spam detection are Support Vector Machine, Random Forest, K-Nearest Neighbours and Decision Tree. The dataset used for this project will consist of spam and ham (normal) emails. The dataset is split into 2 sub-datasets; say

"train dataset" and "test dataset" in the proportion of 75:25. Then the "train" and "test" dataset are then passed as parameters for text-processing. The extracted text is converted into vector for further processing. Then the model is trained for the respective algorithms using the train dataset. Once the model is ready, it is tested on the test dataset. The performance of each algorithm is then analysed using data visualization and performance metrics depicting accuracy of classifying spam and ham emails. This approach helps to conclude which algorithm has the highest accuracy.

Block Diagram:

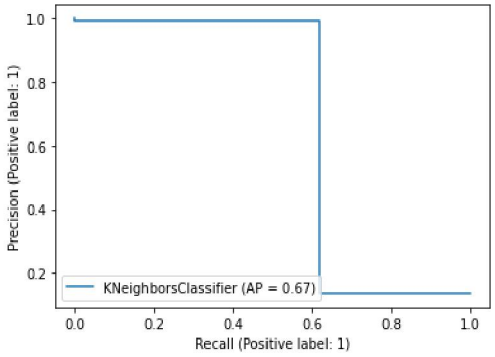


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

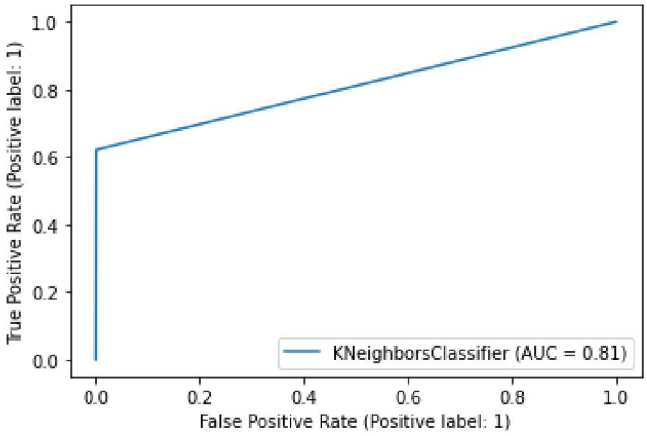
V. ADVANTAGES OF THIS PROJECT

- **Higher Accuracy in inspection is achieved:** Spam Email detection makes use of various machine learning algorithm used for classification which helps in increasing the accuracy of the model. Higher accuracy leads in better classification of spam emails.
- **Manual errors are eliminated:** Our project helps in filtering out spam emails and thereby avoiding manual errors of deleting some important emails in search of spam emails. It helps in providing higher accuracy in filtering out some important emails.
- **Faster results:** Our project not only tries to filter out spam emails but also tries to improve the user experience by providing them results in less amount of time. Various concepts of background processing are used which helps in improving the speed of execution of the app.
- **Limitations/Constraints of Project:**
- **High computational cost:** Spam Email Detection makes use various machine algorithm for classification of emails which requires more time and power for computation.
- **Need a lot of training data:** Machine learning makes use of classification algorithm which requires a lot of training data to train on in order to improve the accuracy of the model and thereby provide better results.

Performance Analysis:
K-nearest neighbour:



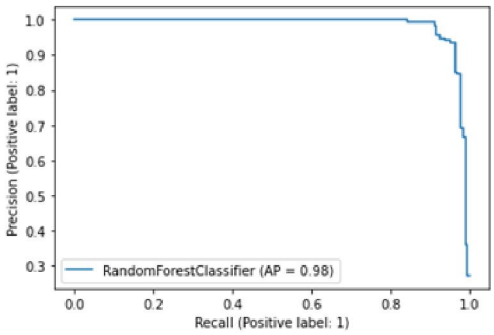
PA1.Classifier.



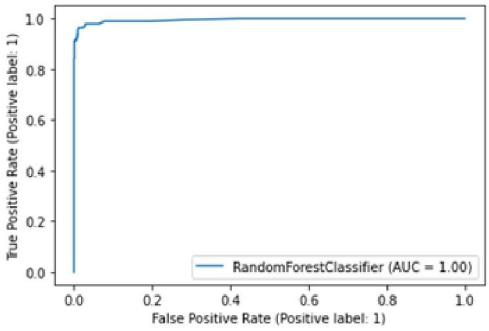
PA2.false rate.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Random Forest:

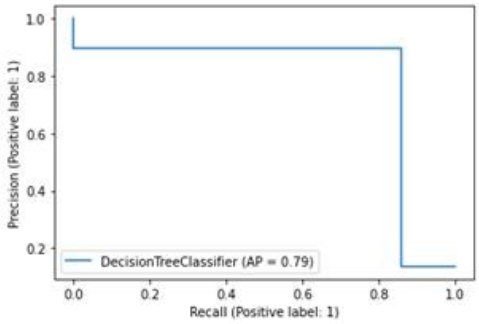


PA3.Classifier

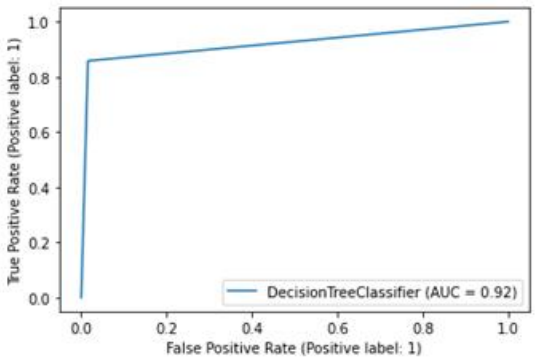


PA4.false rate.

Decision Tree:



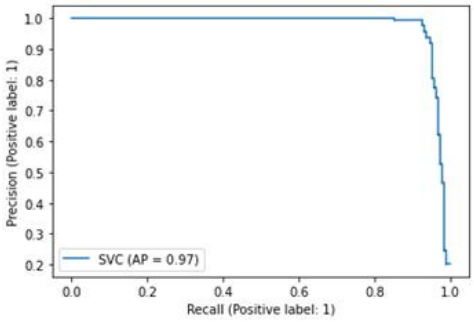
PA5.Classifier.



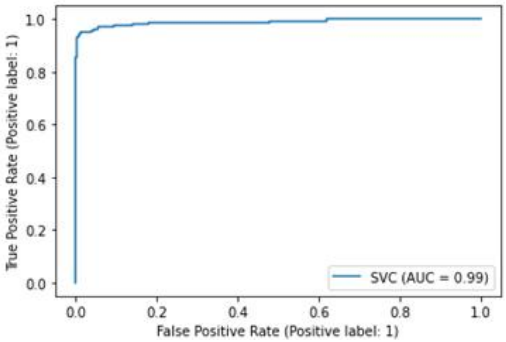
PA6.false rate.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Support Vector Machine(SVM):



PA7.Classifier.



PA8.false rate.

VII. APPLICATIONS

- Email Client Software's: This project can be widely as an email client software where in the user can read the emails and mainly automatically or manually as per user connivence detect the spam emails and delete them. It helps user in securing their personal information from various attacks from emails.
- Job Portals: Email classification project can be used as a utility function in various job portals. It can help in classifying the in-app emails by running the classification in background, thus securing the user's personal information.

VIII. CONCLUSION

Nowadays, Spam is the major concern for email users in the internet. Productivity of Email spam filtering depends on the classification algorithm used. In this paper, Email spam classification using various Machine Learning techniques for the dataset is discussed, concluding the overview of several Spam Filtering techniques and summarizing the accuracy of different proposed approach regarding several parameters. 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.

REFERENCES

[1]. Hu, Y., Zhang, Y., Gong, D. (2021). Multiobjective particle swarm optimization for feature selection with fuzzy cost. IEEE Transactions on Cybernetics, 51(2), 874–888. DOI 10.1109/TCYB.2020.3015756.

[2]. Suryawanshi, Shubhangi&Goswami, Anurag & Patil, Pramod. (2019). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. 69-74.10.1109/IACC48062.2019.8971582.

[3]. NaghmehMoradpoor, Benjamin Clavie, Bill Buchanan,"Employing Machine Learning Techniques for Detection and Classification of Phishing Emails", IEEE Computing Conference, pp 149-156, 2017

[4]. Dr.SwapnaBorde, Utkarsh M. Agrawa,Viraj S. Bilay, Nilesh M. Dogra, "Supervised Machine Learning techniques for Spam Email Detection", IJSART – Vol.3, Issue.3, pp 760-764,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

[5]. 2017. AnjuRadhakrishnan, Vaidhehi V, "Email Classification Using Machine Learning Algorithms", International Journal of Engineering and Technology (IJET), Vol.9, No.2, pp 335340, 2017.

[6]. 4.Mohammad R, Mustafa A. A lifelong spam emails classification model. ApplComput Inform.2020; 18(1/2). doi: 10.1016/j.aci.2020.01.002.

[7]. Sharma S, Amit A. Adaptive approach for spam detection. Inter National J Computer Sci Issues(IJCSI). 2013; 10(4): 23.