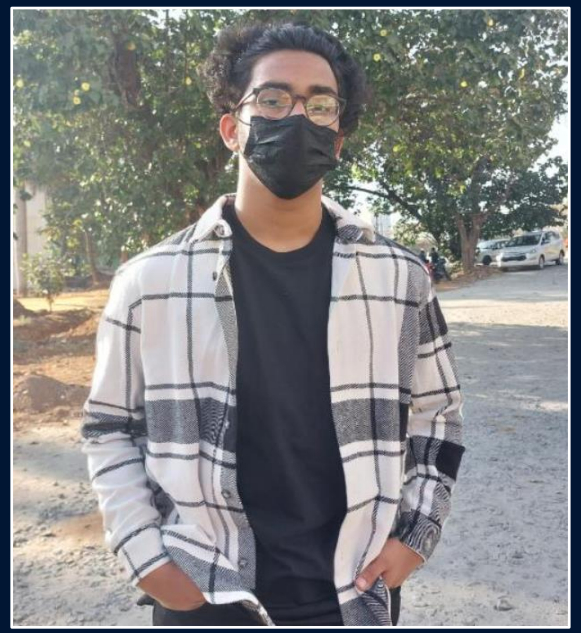
UE20CS302 – MACHINE INTELLIGENCE

Machine Intelligence – MINI Project

Email Spam Detection System

Team Members

Adarsh Kumar
SRN: PES2UG20CS016



Sreekar Govind
SRN: PES2UG20CS125

Aryaman Yadav
SRN: PES1UG20CS079

Problem Statement

In our every day lives, we communicate with various kinds of people regarding many things, especially through the form of ‘Email’. We Send and Receive various kinds of Emails Everyday, regarding Business, Personal mails, or anything really.

But sometimes, we receive Emails that are not beneficial to us, and can be misleading and dangerous, we wouldn’t even know the harm behind them. These Emails come in the form of ‘Spam’.

Spam Emails consist of unimportant, non-beneficial, or even virus embedded messages where it clogs our inbox and we would not be safe with them.

Hence we make this program to filter the various spam messages we get from our Primary/Day-to-Day messages.

Application and Uses

APPLICATION

1. These methods can also be used as ‘AD-Blockers’ to block unwanted AD’s from websites.

2. These methods can be used in a encode and decode system too to filter out encrypted messages.

USES

1. Used to detect unsolicited, unwanted and virus-infected emails and prevent those messages from getting to a user's inbox.

2. It is used in our daily usage of email system where spam emails get filtered from non-spam emails.

Scope and Novelty

Scope

1. The scope of the project is to ensure that this program detects unsolicited, unwanted and virus infected emails and prevent those messages from getting it into a user's inbox.

2. We focus on categorizing the emails into two categories, that is 'SPAM' and 'HAM'(non-spam).

Novelty

1. We come up with a program that derives the best accuracy and precision for using the appropriate model for Filtering Spam Emails.

2. We even made a GUI application to check if the message we pass is SPAM/HAM.

Literature Survey
(3 papers by Student 1)

Title of the paper	Year of Publication	Journal/Conference Name	Advantages	Limitations
Content Based Spam E-mail Filtering Content Based Spam E-mail Filtering	**2016**	International Conference on Collaboration Technologies and Systems (CTS)	Spam filters save time that you could have wasted on removing spam from your Inbox.	Spam filtering is machine-based so there is a room for mistakes called “false positives.”
Existing Spam Filtering Methods Considering different technique	**2021**	International Conference on Technological Advancements and Innovations (ICTAI)	Suspected keywords are known and also with the availability of best heuristic function.	It is difficult to detect or find the suspected IP and also contains errors.
Personalized Classification of Non-Spam Emails Using Machine Learning Techniques	**2022**	International Research Conference on Smart Computing and Systems Engineering (SCSE)	The label Accuracy with the default table is 95%	No one knows how machine will react to sensitive data which will we shared on the mail.

Literature Survey
(3 papers by Student 2)

Title of the paper	Year of Publication	Journal/Conference Name	Advantages	Limitations
A Proposed Data Science Approach for Email Spam Classification using Machine Learning Techniques	**2017**	Internet of Things Business Models, Users, and Networks	A three-tier architecture which is also a client-server architecture is incorporated in the proposed model.	This model blocks the email of the sender who are likely to spam from a predefined list by the system administrator
A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms	**2021**	Mansoor RAZA and Nathali Dilshani Jayasinghe School of Computing and Mathematics Charles Sturt University, Study Centre Melbourne VIC 3000, Australia	Good Efficiency, Greater accuracy	Naive Bayes is one of the utmost well-known algorithms applied in these procedures, rejecting sends essentially dependent on content examination can be a difficult issue in the event of bogus positives.
Cascaded Simple Filters for Accurate and Lightweight Email-Spam Detection	**2010**	2010 Fourth International Conference on Emerging Security Information, Systems and Technologies	Sensitivity, specificity and accuracy	Low performance.

Literature Survey
(3 papers by Student 3)

Title of the paper	Year of Publication	Journal/Conference Name	Advantages	Limitations
Spam filtering email classification (SFECM) using gain and graph mining algorithm	**2017**	2017 2nd International Conference on Anti-Cyber Crimes (ICACC)	Accuracy, precision and recall	Very low performance.
Artificial Intelligence-Based Methods For Filtering Spam Messages In Email Services	**2021**	2021 International Conference on Information Science and Communications Technologies (ICISCT)	Accuracy, precision, recall, and F-measure.	Time taken to build MLP is very high. No improvement on existing methods.
Hybrid Spam E-mail Filtering	**2009**	2009 First International Conference on Computational Intelligence, Communication Systems and Networks	Global best positions.	Standard evaluation metrics were not used to evaluate the performance of the proposed method.

Proposed Approach

We first import our dataset that consists of Spam **&** *Non- Spam Emails. We then import the necessary libraries to perform certain operations on our dataset.*

We perform a Basic Data Analysis where we display the dataset we have and we clean our data where null/duplicate **&** *redundant values are removed.*

We perform a EDA[Exploratory Data Analysis] to check the percentage of Spam and Non-Spam Emails we make a top **50** *Spam used words/list and store it under a target data-frame to use for training.*

We then split the data to use for training and testing and perform the training using various models [such as Naïve Bayes, SVM, etc....] and we then compare the precision and accuracy to get the Best Fit Model.

Results and Discussion

High adoption rate for supervised machine learning approach can be seen throughout the review. This approach is used mainly because it generates higher accuracy results with less variation giving high consistency for this approach.

Apart from that, we have found out that certain algorithms such as Naiive Based and SVM have high demand compared to other Machine Learning Algorithms.

The multi algorithm used systems are more common in use to cater better outcome rather than using single algorithm.

Researchers have more focused on email features such as Bow and Body text creating future research opportunities to develop systems to detect spam on other email features

References
(6 Literature Survey Papers web links)

[1] *L. Firte, C. Lemnaru, R. Potolea, "Spam Detection Filter using KNN Algorithm and Resampling," Intelligent Computer Communication and Processing (ICCP),* 2010 *IEEE International Conference on, pp.* 27–33, 26-28 *Aug* 2010.

[2] *Drucker et al.*(1999), *for classifying the email text content used (support vector machines method, then compared it with other algorithm like Boosting decision trees , Ripper & Racchio* [21]

[3] *S. P. Gautam, "Email classification using a self-learning technique based on user preferences", Master. dissertation, Dept. Comp. Sc., North*

[4] *Yoo, S.*(2010), *Machine learning methods for personalized email prioritization. PhD. Carnegie Mellon University*

[5] *F. Qian, Y. C. H. Abhinav Pathak, Z. M. Mao, and Y. Xie, "A case for unsupervised-learning-based spam filtering," Univ. Minnesota J.*, 2010.

[6] *An Efficient Spam Filtering Techniques for Email Account S. Roy, A. Patra, S.Sau, K.Mandal, S. Kunar*

Thank You