International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.108 ISSN: 2581-4621

Email Spam Detection Using Machine Learning Algorithms

Labbe shaik Ahmed[1], K.Samson Paul[2]
[1]*MCA Student, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*
[2]*Assistant Professor, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh, India*

Abstract
Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. S ending malicious link through spam emails which can harm our system and can also seek in into your system. Creating a fake profile and email account is much easy for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds. S o, it is needed to Identify those spam mails which are fraud, this project will identify those spam by using techniques of machine learning, this paper will discuss the machine learning algorithms and apply all these algorithm on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy.

Introduction
Email or electronic mail spam refers to the “using of email to send unsolicited emails or advertising emails to a group of recipients. Unsolicited emails mean the recipient has not granted permission for receiving those emails. “The popularity of using spam emails is increasing since last decade. Spam has become a big misfortune on the internet. Spam is a waste of storage, time and message speed. Automatic email filtering may be the most effective method of detecting spam but nowadays spammers can easily bypass all these spam filtering applications easily. Several years ago, mos t of the spam can be blocked manually coming from certain email addresses. Machine learning approach will be used for spam detection. Major approaches adopted closer to junk mail filtering encompass “text analysis, white and blacklists of domain names, and community-primarily based techniques”. Text assessment of contents of mails is an extensively used method to the spams. Many answers deployable on server and purchaser aspects are available. Naive Bayes is one of the utmost well-known algorithms applied in these procedures. However, rejecting sends essentially dependent on content examination can be a difficult issue in the event of bogus positives. Regularly clients and organizations would not need any legitimate messages to be lost. The boycott approach has been probably the soonest technique pursued for the separating of spams. The technique is to acknowledge all the sends other than those from the area/electronic mail ids. Expressly boycotted. With more up to date areas coming into the classification of spamming space names this technique keeps an eye on no longer work so well. The white list approach is the approach of accepting the mails from the domain names/addresses openly whitelisted and place others in a much less importance queue, that is delivered most effectively after the sender responds to an affirmation request sent through the “junk mail filtering system”

Literature Survey
A Comprehensive Survey for Intelligent Spam Email Detection.
AUTHOR: Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M.
Abstract:
The tremendously growing problem of phishing e-mail, also known as spam including spear phishing or spam borne malware, has demanded a need for reliable intelligent anti-spam e-mail filters. This survey paper describes a focused literature survey of Artificial Intelligence (AI) and

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.108 ISSN: 2581-4621

Machine Learning (ML) methods for intelligent spam email detection, which we believe can help in developing appropriate countermeasures. In this paper, we considered 4 parts in the email's structure that can be used for intelligent analysis: (A) Headers Provide Routing Information, contain mail transfer agents (MTA) that provide information like email and IP address of each sender and recipient of where the email originated and what stopovers, and final destination. (B) The SMTP Envelope, containing mail exchangers' identification, originating source and destination domains\users. (C) First part of SMTP Data, containing information like from, to, date, subject - appearing in most email clients (D) Second part of SMTP Data, containing email body including text content, and attachment. Based on the number the relevance of an emerging intelligent method, papers representing each method were identified, read, and summarized. Insightful findings, challenges and research problems are disclosed in this paper. This comprehensive survey paves the way for future research endeavors addressing theoretical and empirical aspects related to intelligent spam email detection

Existing System

okenization is the process of splitting a stream of manuscript into phrase, symbols, words, or any expressive elements named as tokens." The rundown of token further utilized for contribution for additional handling, for example, content mining and parsing. Tokenization is valuable in both semantics (where it is as content division), and as lexical examination in software engineering and building. It is occasionally hard to define what is intended by the term "word". As tokenization happens at the word level. Frequently a token trusts on modest heuristics, for instance: Tokens are parted by whitespaces characters, like "line break" or "space", or by "punctuation characters". Every single neighboring string of alphabetic characters are a piece of one token; similarly, with numbers. White spaces and punctuations might or might not involve in the resulting lists of tokens

DISADVANTAGES OF EXISTING SYSTEM :

1) Less accuracy
2)low Efficiency

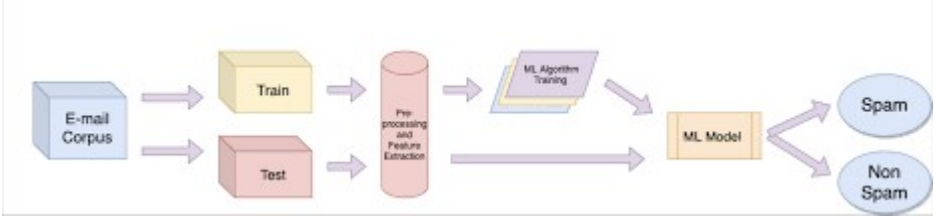
Proposed System

When the data is considered, always a very large data sets with large no. of rows and columns will be noted. But it is not always the case the data could be in many forms such as Images, Audio and Video files Structured tables etc. Machine doesn't understand images or video, text data as it is, Machine only understand 1s and 0s. Steps in Data Preprocessing: Data cleaning: In this step the work like filling of "missing values", "smoothing of noisy data", "identifying or removing outliers ", and "resolving of inconsistencies is done." Data Integration: In this step addition of several databases, information files or information set is performed. Data transformation: Aggregation and normalization is performed to scale to a specific value Data reduction: This section obtains a summary of the dataset which is very

ADVANTAGES OF PROPOSED SYSTEM :

1) High accuracy
2)High efficiency

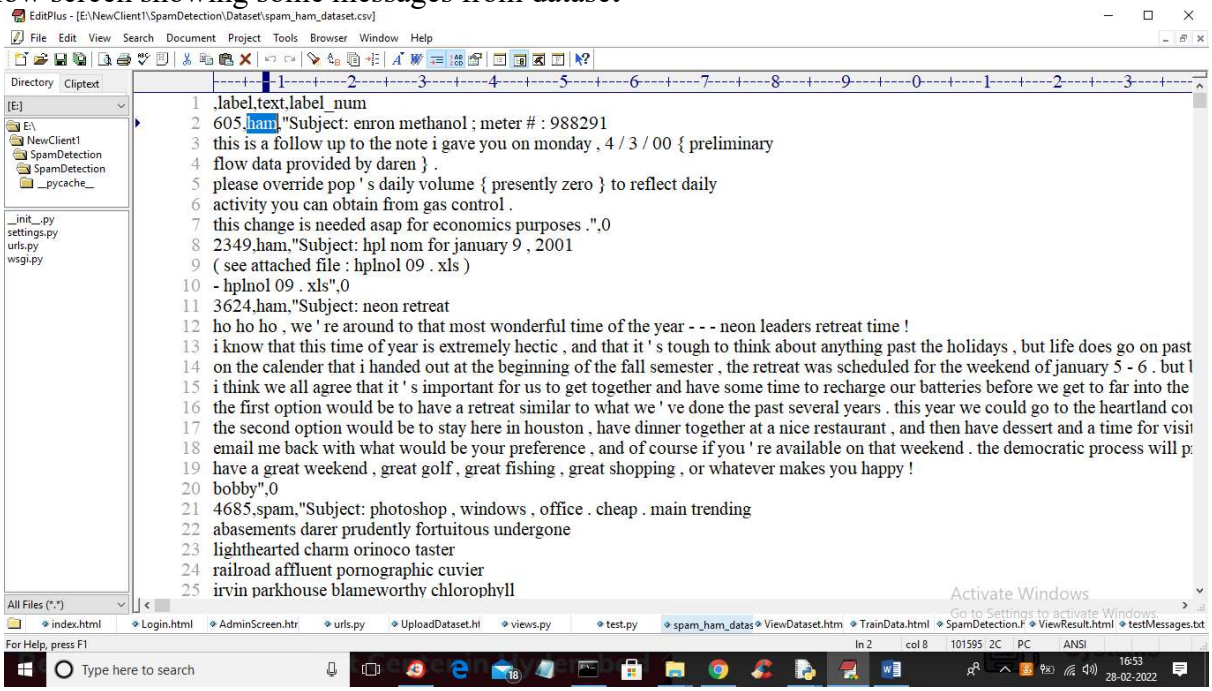
System Architecture



International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.108 ISSN: 2581-4621

Results

In this project we are using Random Forest machine learning algorithm to predict EMAIL messages as SPAM or HAM. To train random forest we have used KAGGLE EMAIL SPAM dataset and below screen showing some messages from dataset



In above screen first row represents dataset column names and remaining rows contains EMAIL message and class label as HAM or SPAM and by using above dataset we will train Random Forest algorithm. After training we can input any message then random forest will predict as SPAM or HAM.

To implement this project we have designed following Modules

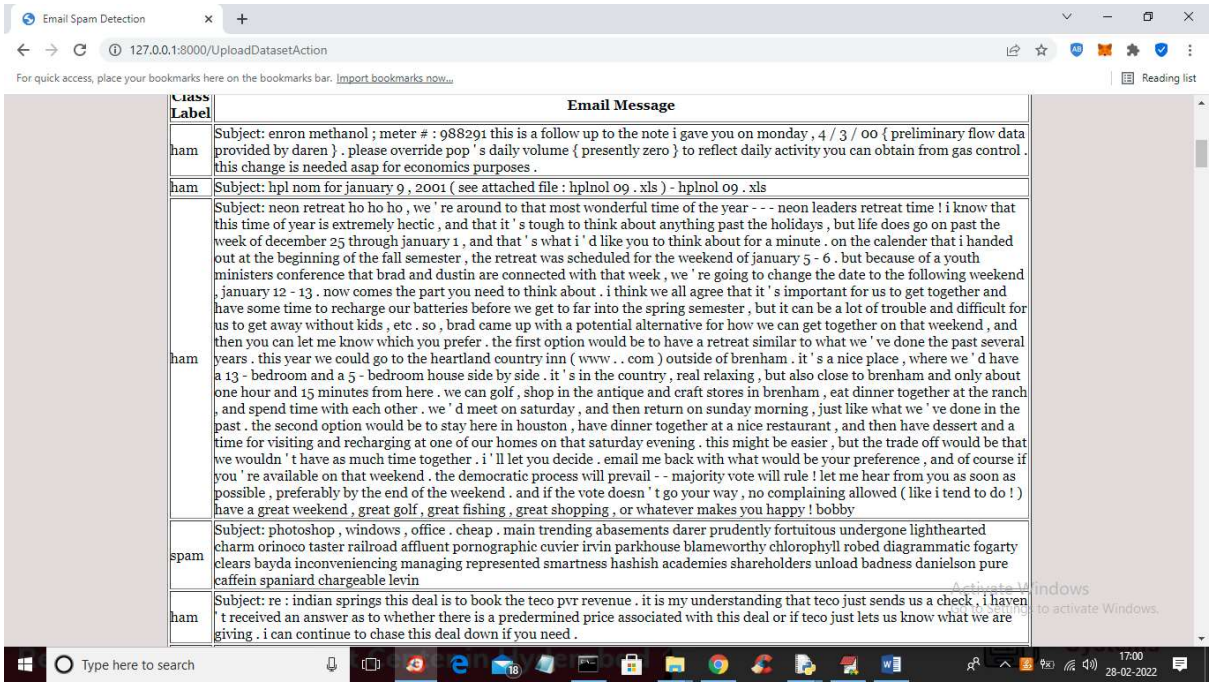
1)Admin Login: using this module admin can login to application by using username and password as ‘admin’ and ‘admin’

2)Upload Dataset: after login admin can click on ‘Upload Dataset’ link to upload dataset to application

3)Train Dataset Using Random Forest: now admin can click on this link to train random forest algorithm and this model will be applied on TEST data to calculate random forest prediction accuracy.

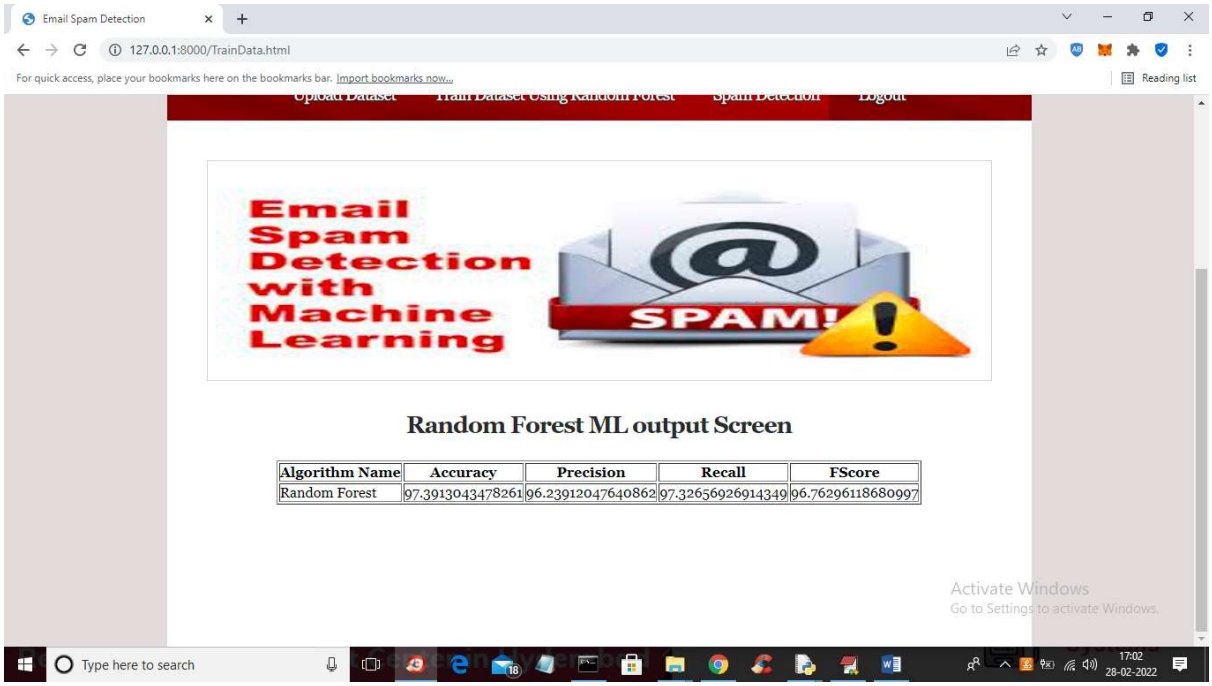
4)Spam Detection: admin can click on this link and then enter some message and then press submit button and then Random Forest will predict that message as HAM or SPAM

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.108 ISSN: 2581-4621



Class Label	Email Message
ham	Subject: enron methanol ; meter # : 988291 this is a follow up to the note i gave you on monday , 4 / 3 / 00 { preliminary flow data provided by daren } . please override pop ' s daily volume { presently zero } to reflect daily activity you can obtain from gas control . this change is needed asap for economics purposes .
ham	Subject: hpl nom for january 9 , 2001 (see attached file : hplnol 09 . xls) - hplnol 09 . xls
ham	Subject: neon retreat ho ho ho , we ' re around to that most wonderful time of the year - - - neon leaders retreat time ! i know that this time of year is extremely hectic , and that it ' s tough to think about anything past the holidays , but life does go on past the week of december 25 through january 1 , and that ' s what i ' d like you to think about for a minute . on the calender that i handed out at the beginning of the fall semester , the retreat was scheduled for the weekend of january 5 - 6 . but because of a youth ministers conference that brad and dustin are connected with that week , we ' re going to change the date to the following weekend , january 12 - 13 . now comes the part you need to think about . i think we all agree that it ' s important for us to get together and have some time to recharge our batteries before we get to far into the spring semester , but it can be a lot of trouble and difficult for us to get away without kids , etc . so , brad came up with a potential alternative for how we can get together on that weekend , and then you can let me know which you prefer . the first option would be to have a retreat similar to what we ' ve done the past several years . this year we could go to the heartland country inn (www . . com) outside of brenham . it ' s a nice place , where we ' d have a 13 - bedroom and a 5 - bedroom house side by side . it ' s in the country , real relaxing , but also close to brenham and only about one hour and 15 minutes from here . we can golf , shop in the antique and craft stores in brenham , eat dinner together at the ranch , and spend time with each other . we ' d meet on saturday , and then return on sunday morning , just like what we ' ve done in the past . the second option would be to stay here in houston , have dinner together at a nice restaurant , and then have dessert and a time for visiting and recharging at one of our homes on that saturday evening . this might be easier , but the trade off would be that we wouldn ' t have as much time together . i ' ll let you decide . email me back with what would be your preference , and of course if you ' re available on that weekend . the democratic process will prevail - - majority vote will rule ! let me hear from you as soon as possible , preferably by the end of the weekend . and if the vote doesn ' t go your way , no complaining allowed (like i tend to do !) have a great weekend , great golf , great fishing , great shopping , or whatever makes you happy ! bobby
spam	Subject: photoshop , windows , office . cheap . main trending abasements darer prudently fortuitous undergone lighthearted charm orinoco taster railroad affluent pornographic cuvier irvin parkhouse blameworthy chlorophyll robed diagrammatic fogarty clears bayda inconveniencing managing represented smartness hashish academies shareholders unload badness danielson pure caffein spaniard chargeable levin
ham	Subject: re : indian springs this deal is to book the teco pvr revenue . it is my understanding that teco just sends us a check , i haven ' t received an answer as to whether there is a predermined price associated with this deal or if teco just lets us know what we are giving . i can continue to chase this deal down if you need .

In above screen dataset loaded and we can see class label and email messages and now click on 'Train Dataset Using Random Forest' link to train random forest and get below output

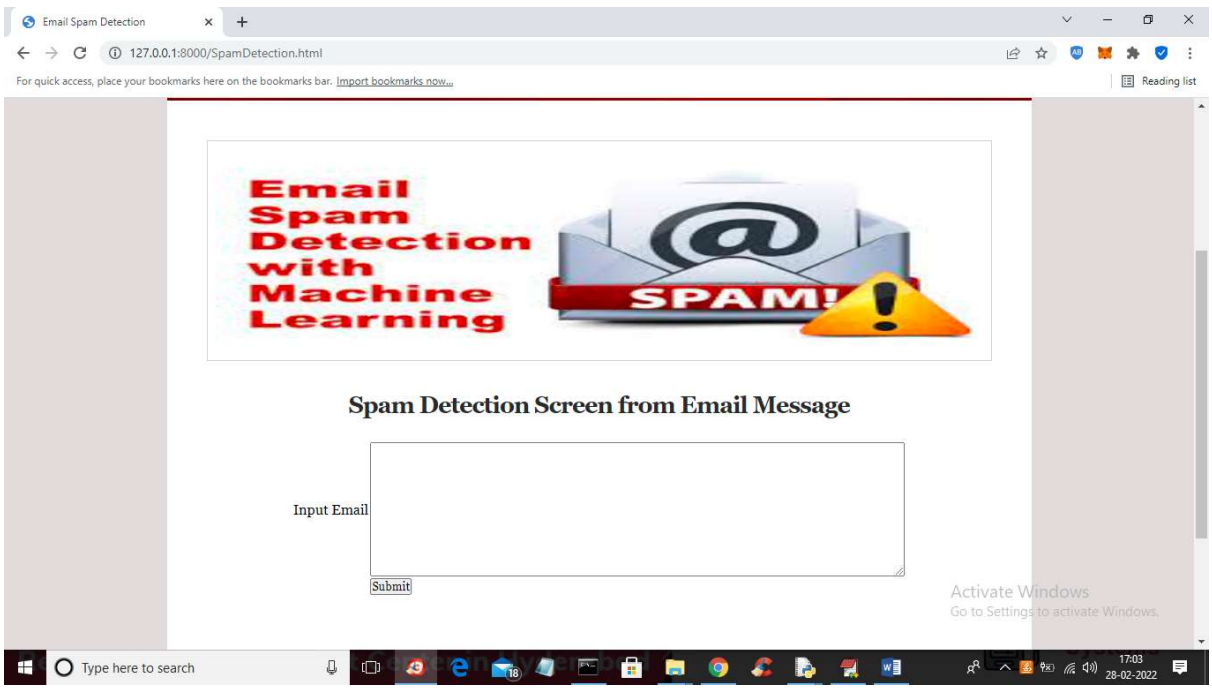


Random Forest ML output Screen

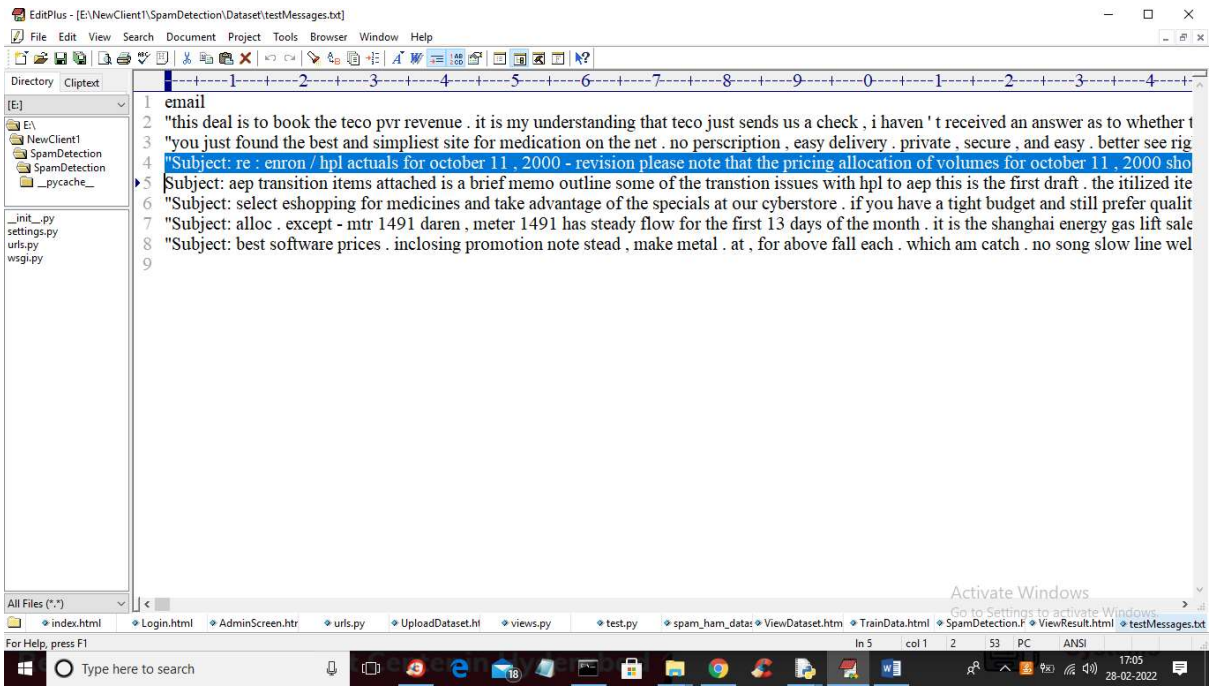
Algorithm Name	Accuracy	Precision	Recall	FScore
Random Forest	97.39130434782612	96.23912047640862	97.32656926914349	96.76296118680997

In above screen Random Forest trained and we got its prediction accuracy as 97% and we can see precision, recall and FSCORE. Now random forest is trained and now click on 'Spam Detection' link to get below screen

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.108 ISSN: 2581-4621

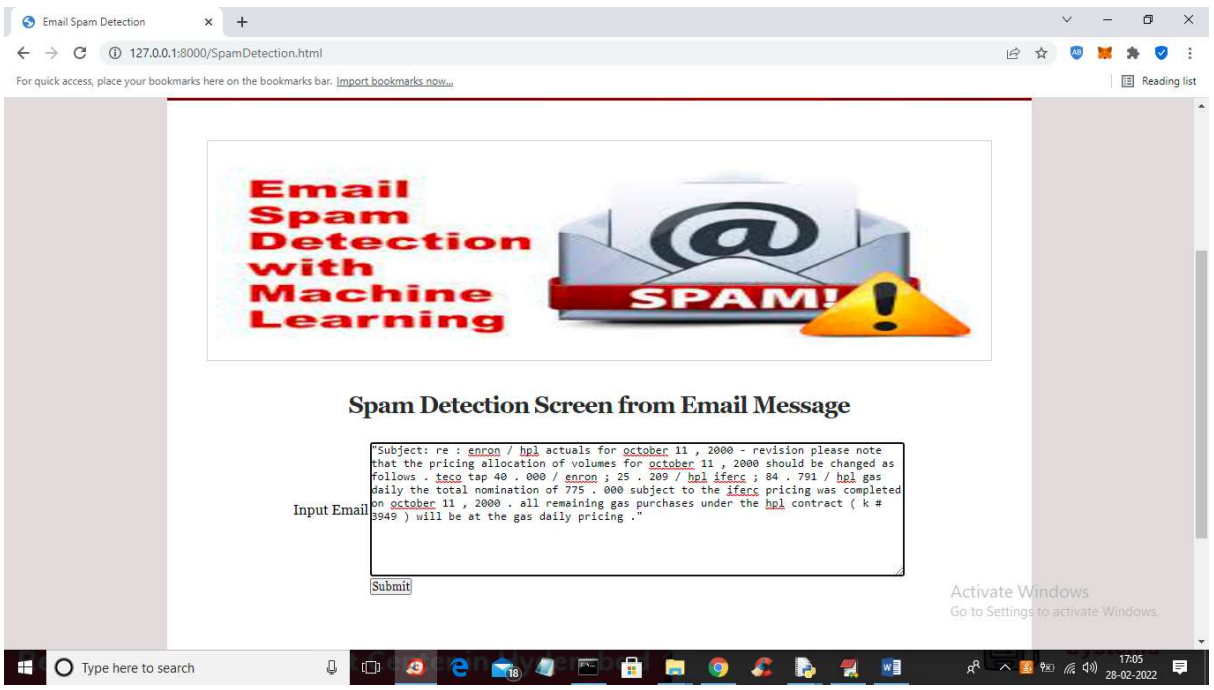


In above screen you can input some message and if you don't know what to enter then you can copy some text from 'testMessages.txt' from 'Dataset' folder and paste in above screen

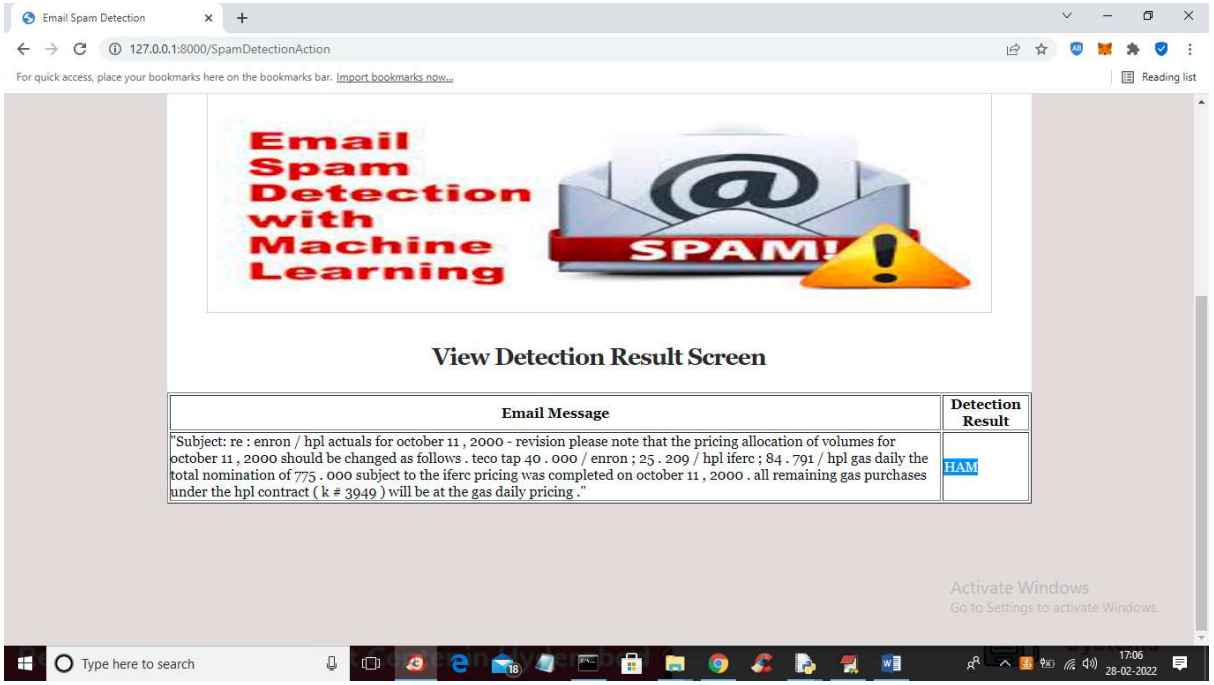


In above screen I am copying one line of text and paste in text area like below screen

International Journal of Engineering Technology and Management Science
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.108 ISSN: 2581-4621



In above screen I pasted the message and press submit button to get below output



In above screen in blue colour text we can see message is detected as HAM and similarly you can paste other messages and get result

Conclusion

With this result, it can be concluded that the Multinomial Naïve Bayes gives the best outcome but has limitation due to class-conditional independence which makes the machine to misclassify some tuples. Ensemble methods on the other hand proven to be useful as they using multiple classifiers for class prediction. Nowadays, lots of emails are sent and received and it is difficult as our project is only able to test emails using a limited amount of corpus. Our project, thus spam detection is proficient of filtering mails giving to the content of the email and not according to the domain names or any other criteria. Therefore, at this it is an only limited body of the email. There is a wide

DOI:10.46647/ijetms.2025.v09i02.108 ISSN: 2581-4621

possibility of improvement in our project. The subsequent improvements can be done: “Filtering of spams can be done on the basis of the trusted and verified domain names.” “The spam email classification is very significant in categorizing e-mails and to distinct e-mails that are spam or non-spam.” “This method can be used by the big body to differentiate decent mails that are only the emails they wish to obtain.”

REFERENCES:

1. 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.
2. Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access, 7, 168261-168295. [08907831]. https://doi.org/10.1109/ACCESS.2019.2954791
3. K. Agarwal and T . Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.
4. Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "T ext and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty, and Information T echnology (ICROIT), 2014 International Conference on, pp.153 -155. IEEE, 2014
5. Mohamad, Masurah, and Ali Selamat. "An evaluation on t he efficiency of hybrid feature selection in spam email classification." In Computer, Communications, and Control T echnology (I4CT), 2015 International Conference on, pp. 227 -231. IEEE, 2015
6. Shradhanjali, Prof. T oran Verma “E-Mail Spam Detection and Classification Using SVM and Feature Extraction”in International Jouranl Of Advance Reasearch, Ideas and Innovation In T echnology,2017 ISSN: 2454-132X Impact factor: 4.295
7. W.A, Awad & S.M, ELseuofi. (2011). Machine Learning Methods for Spam E-Mail Classification. International Journal of Computer Science & Information Technology. 3. 10.5121/ijcsit.2011.3112.
8. A. K. Ameen and B. Kaya, "Spam detection in online social networks by deep learning," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, T urkey, 2018, pp. 1-4.
9. Diren, D.D., Boran, S., Selvi, I.H., & Hatipoglu, T . (2019). Root Cause Detection with an Ensemble Machine Learning Approach in the Multivariate Manufacturing Process.
10. T asnim Kabir, Abida Sanjana Shemonti, Atif Hasan Rahman. "Notice of Violation of IEEE Publication Principles: Species Identification Using Partial DNA Sequence: A Machine Learning Approach”, 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), 2018