A Mini Project Synopsis on

E-Mail Spam Detection

T.E. - I.T Engineering

Submitted By

Sameer Sawant	**19104054**
Raj Shisode	**19104070**
Siddhart Chhoriya	**20204005**

Under The Guidance Of

Prof. Apeksha Mohite

DEPARTMENT OF INFORMATION TECHNOLOGY

A.P.SHAH INSTITUTE OF TECHNOLOGY

G.B. Road, Kasarvadavali, Thane (W), Mumbai-400615

UNIVERSITY OF MUMBAI

Academic year : 2021-22

CERTIFICATE

This to certify that the Mini Project report on E-Mail Spam Detection has been submitted by Sameer Sawant(19104054),Raj Shisode(19104070) and Siddhart Chhoriya (20204005) who are a Bonafide students of A. P. Shah Institute of Technology, Thane, Mumbai, as a partial fulfilment of the requirement for the degree in **Information Technology**, during the academic year **2021-2022** in the satisfactory manner as per the curriculum laid down by University of Mumbai.

Ms Apeksha Mohite
Guide
Ms Roshni Singh
Co -Guide

Prof. Kiran Deshpande
Head Department of Information Technology

Dr. Uttam D.Kolekar
Principal

External Examiner(s)
1.
2.

Place:A.P.Shah Institute of Technology, Thane
Date:

ACKNOWLEDGEMENT

This project would not have come to fruition without the invaluable help of our guide **Prof. Apeksha Mohite** and **Prof. Roshni Singh.** Expressing gratitude towards our HOD, **Prof. Kiran Deshpande**, and the Department of Information Technology for providing us with the opportunity as well as the support required to pursue this project. We would also like to thank our teacher **Prof. Nahid Shaikh** who gave us her valuable suggestions and ideas when we were in need of them. We would also like to thank our peers for their helpful suggestions.

TABLE OF CONTENTS

Chapter 1

Introduction

Apart from the numerous benefits and conveniences people around the world one can enjoy due to the Internet, there are also multiple drawbacks. Not all of them are obvious to an average user, and perhaps only professional IT workers face them from time to time. However, there is a problem almost every Internet user has encountered at least once in a lifetime. Unlike many people might think, spam is not just an annoying email message; in fact, spam can be a dangerous tool capable of harming its recipients, and should be outlawed.

Even though spam messages usually have an "unsubscribe" link, getting off a spammer's list requires a number of actions, such as visiting the website, acknowledging un subscription option, sending confirmation letters, typing the captcha, and so on. This might be not a problem in the case of being a target of several spammers; however, usually Internet users receive dozens of spam messages daily; unsubscribing from each of them is almost impossible.

1.1. Purpose:

Implementing spam filtering is extremely important for any organization. Spam emails, also known as junk email involves nearly identical messages sent to numerous recipients by email.

Not only does spam filtering help keep garbage out of email inboxes, it helps with the quality of life of business emails because they run smoothly and are only used for their desired purpose. Spam filtering is essentially an anti-malware tool, as many attacks through email are trying to trick users to click on a malicious attachment, asking them to supply their credentials, and much more.

1.2 Objectives:

- To create an Email Spam Detector, it should be able Detect content of the given message instantaneously without any delay and provide at most accurate or answer which the users were looking for.
- To alert the user about suspicious emails that are sent by unauthorized users. To identify whether an email is a spam or a ham.
- To learn about the use of machine learning algorithm in E mail spam detection process.

1.3 Scope:

The project provides sensitivity to the client and helps to adapt to the future spam techniques. It also considers a complete message instead of single words with respect to its content. It reduces the IT administration workload as now they will have to deal with less important emails and more of important emails that require urgent attention. Email Spam detection method can also be used to reduce cyber criminal activities because the email spam detector may also filter out those messages that contain malicious links that may automatically download dangerous software in your system or device.

Chapter 2

Problem Definition

Junk emails waste a lot of time and effort that could've been used for something more productive, but that's not even the worst part. Spam is also a popular means of transferring harmful malware and electronic viruses. And in an age where hacking tools and techniques grow more and more sophisticated by the minute, spam-instigated security attacks become a perpetual threat.

Spam emails are also an avenue for marketers to exploit your data's privacy. Responding to just one unsolicited email could put you in the mailing lists of many other companies. Before you know it, your spam emails would've already multiplied tenfold.

Phishing scams have also gained attraction through spamming. A spam email that was made to look like it came from a legitimate entity that you trust (like your bank or someone you know) could end up stealing sensitive information if you're not careful. You could be a victim of identity theft, or you could lose all your money if you by mistakenly hand in your bank details.

Chapter 3

Proposed System

E-mail Spam detector will use Machine Learning technique like Natural Language processing to understand the user input and will predict output as per the trained model and response as per the best matching reply it finds.

In proposed System we will be having an interface wherein we would just have to copy and paste the email message after few seconds of processing data with the help of naïve Bayes algorithm, the system will come up with a message showing whether an email is a spam or a ham.

3.2Features and Functionality

1. Give response instantly – The Email Spam detector is able to provide response on the user input as soon as possible.

2. Easy to understand User Interface – The Graphic User Interface on the email spam predictor should be clean and easy to understand able and have simple display window

3. Accurate response – An Email spam detection system should be able to predict the response as accurate as possible to give an output whether an email is a spam or a ham.

Chapter 4.

Literature Survey:

Title	Author/s	Findings
Study of Spam E mails	Mr Raj Patil, Prof Justice Kumar, Mr. Ashish Tiwary	The author has worked with different machine learning algorithms for email classification such as Neural Network (NN), Support Vector Machine (SVM), J48 Decision Tree based classifier, Naïve Bayes. The dataset used by the author was Spam Base dataset. In this paper work, the author didn't mention advantages and disadvantages of any algorithm.
A systematic literature review on spam content detection and classification .	Mr.Abbas.M and Mr.Ibrahim.M (2020)	Based on our research objective, the initial search keywords were carefully chosen. Following an initial search, new words discovered in several related articles were used to generate several keywords. These keywords were later trimmed to fit the research's objectives. We chose certain search keywords based on the goal of our survey work, and after performing an initial search on those words, several keywords were derived from selected articles.

Comprehensive Survey for Intelligent Spam Email Detection	Mr.Kalaivani.K.M, Dharinishree.K.A, Gowsalya.B, Punitha.P, Prof. Kalaikumaran.T	This research initiative is to address a gap that has risen over time in the field of spam email detection. The current solutions are mostly lagging behind the innovativeness the spammers are constantly bringing in, which heavily justifies the emergence of machine learning based anti-spam propositions
A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques	M.s Shridevi.A.Mali, Sravanthi.G, Prof. Siva Subbha Rao.P, Rajja.S, Sushma. Mr. S.J, Reddy.A.R	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.

Chapter 5.

Project Outcomes

- The outcome of our project is to provide a system where it should be able to respond with correct and accurate response after it process the Information provided from the given input and notify the user with an output whether an email is a spam or not.

- In this system to solve the problem of spam emails, the spam classification system is created to identify spam and non-spam.

- Since spammers may send spam messages many times, it is difficult to identify it every time manually. So this project will be using the algorithm in our proposed system to detect the spam.

Chapter 6

Software Requirements

Software Used –

- Python (3.6 and above),

- Jupyter Notebook,

- Pycharm,

- NLTK (Natural Language Tool Kit),

- Numpy,

- Streamlit,

- Sklearn,

Chapter 7:

Project Design

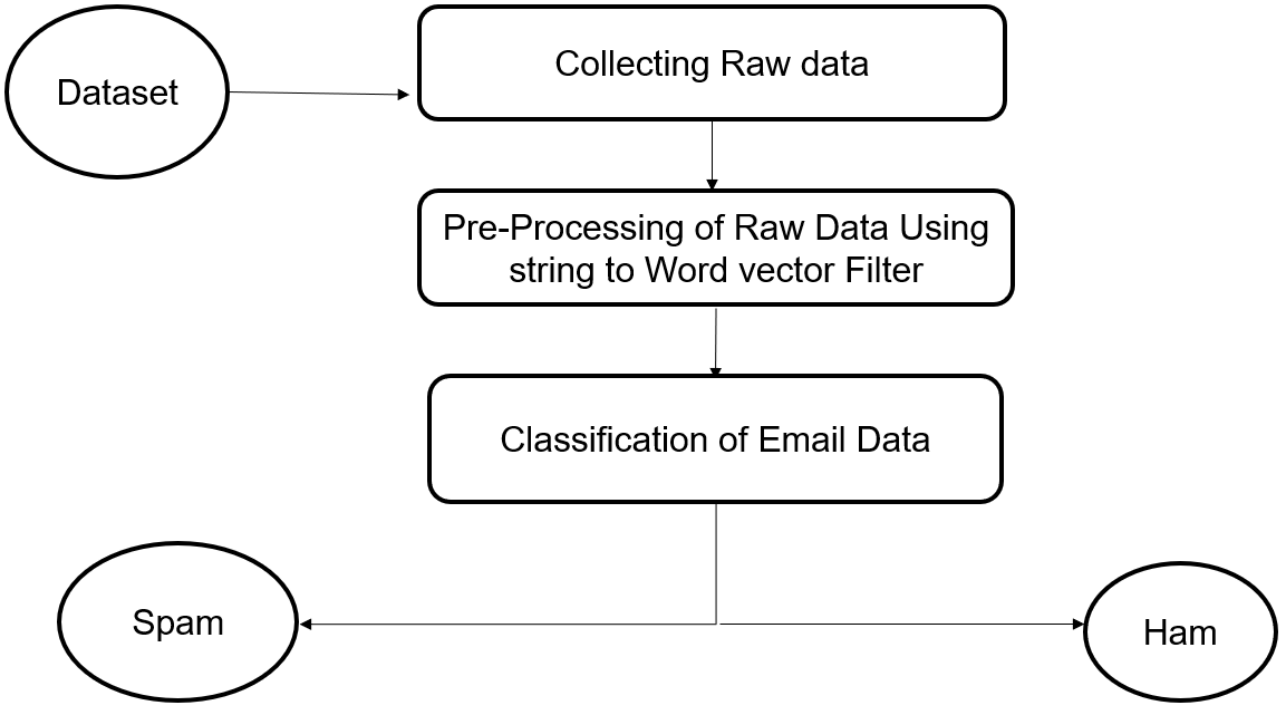


Figure 7.1 : Flow of Modules

For our project the main machine learning code is done using python, Jupyter Notebook, pycharm, numpy and nltk libraries. The flow of modules in our project is that at first we have to train the data which is available at Kaggle, which consists of spam and ham email examples which is used to predict emails. As for training the data first nltk tokenizes the tags, pattern and response and forms vector array and then stemming of porter type is done to the words, converting to lower case, removing special characters, stop words and punctuations is done for data pre-processing so that the words should have basic words meaning rather than being complex. So after all this words are made into bag of words so that words are converted into number using count and tfidf vectorizer as the models are not able to understand words.

Model is made on Jupyter Notebook using in-built modules so the processing of data gets as accurate result as possible with least loss possible error in the prediction. Next process is to connect to the user interface(GUI) which is made using Streamlit. Vectorization and sklearn models are used to divide and process the train and test the data. Voting and stacking is performed to test accuracy and precision of each models and combination of models. The website asks for input mail message, after pressing the predict button the website predicts if the email is spam email or ham email and displays it as the output.

Chapter 8:

Screenshot of Website

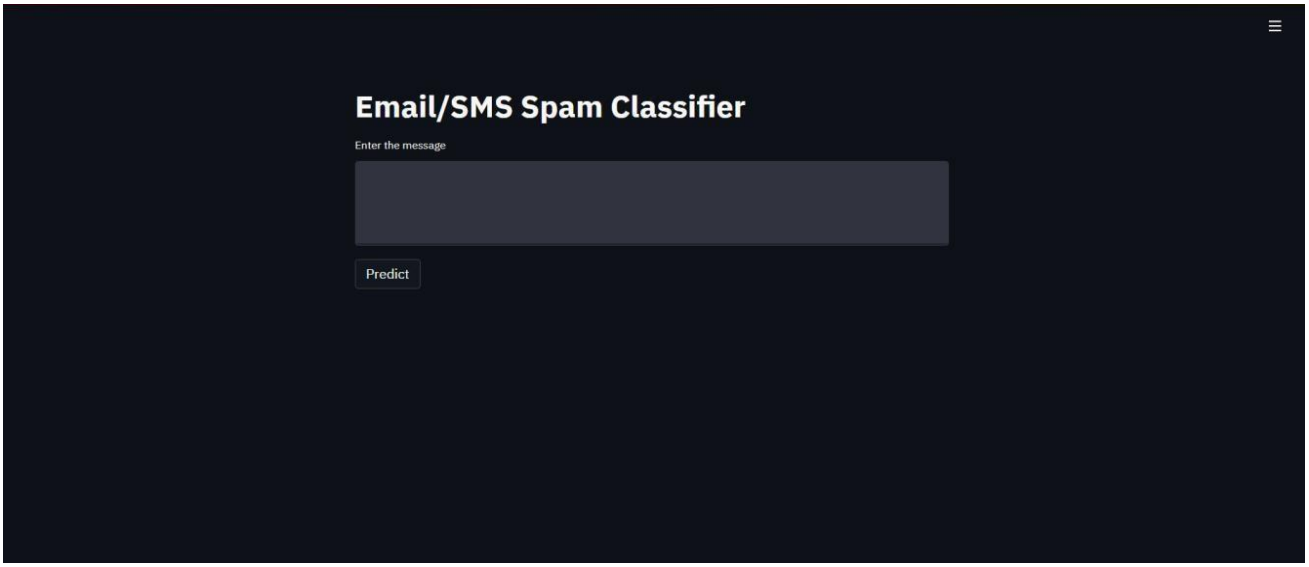


Figure 8.1. Main Page

The above figure shows the main page of our project. We have provided a text area to input the email content. We also have provided a predict button to predict the email as spam or ham.

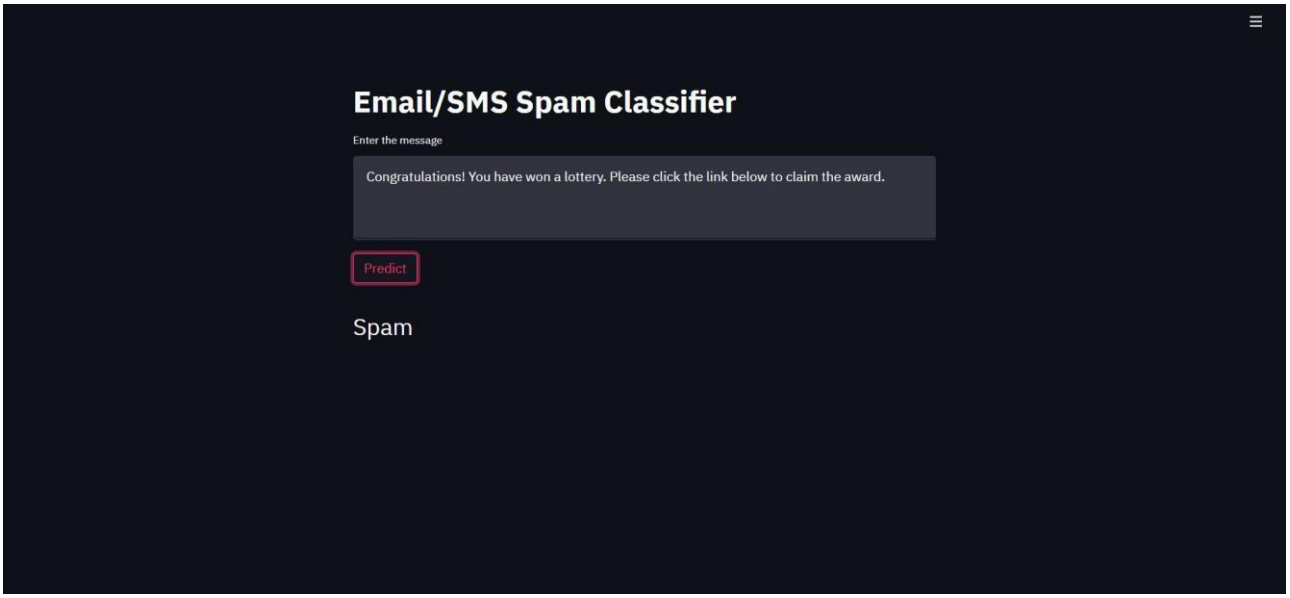


Figure 8.2. Email Detected Spam.

The above screenshot represents the output of the system after we enter a message and click on the predict button, which shows the output as spam.

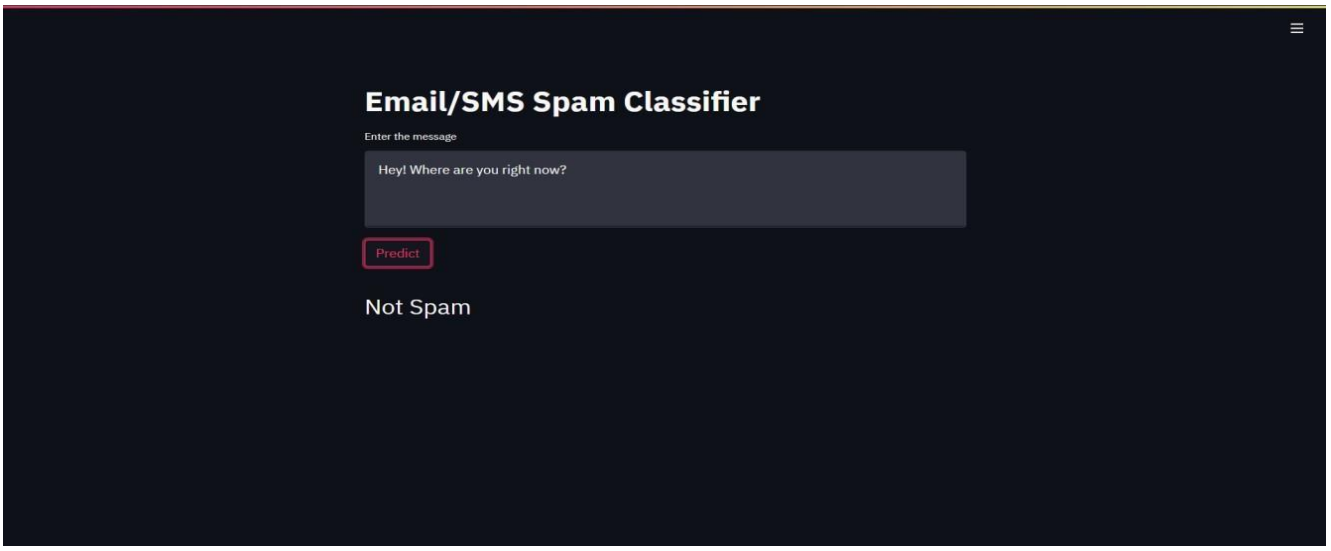


Figure8.3: Email detected not spam.

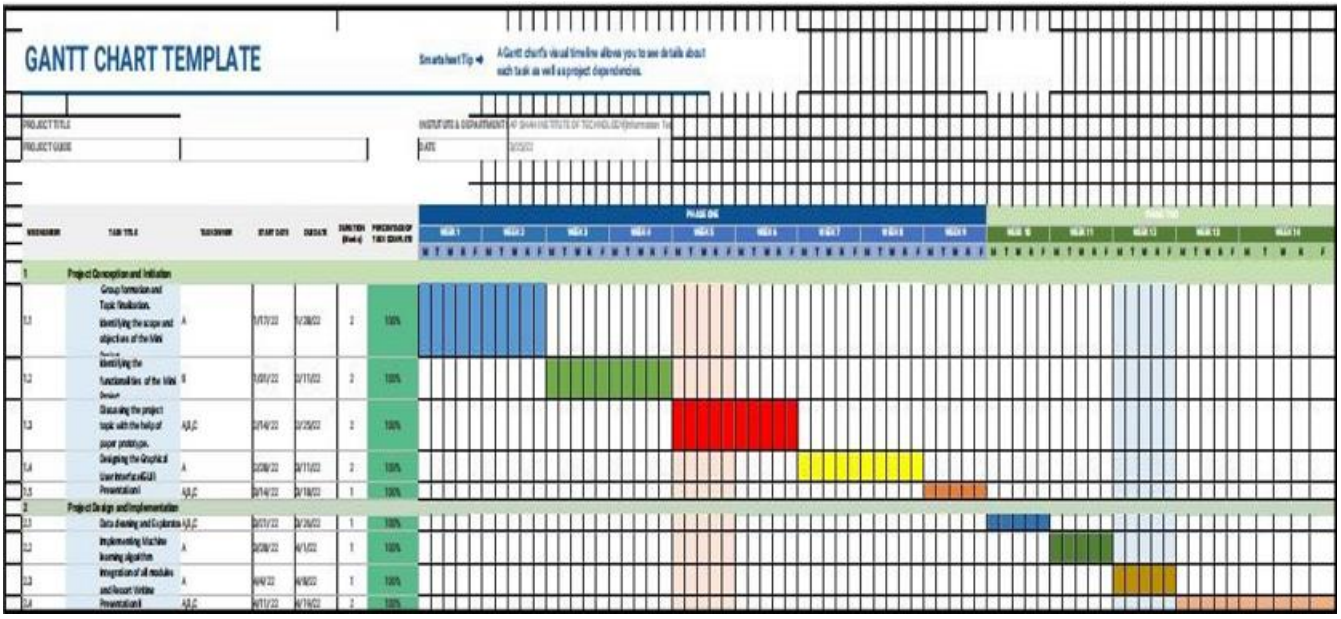
The above screenshot represents the output of the system after we enter a message and click on the predict button, which shows the output as not spam.

Chapter 9:

Project Scheduling Template

Sr. No	Group Member	Time duration	Work to be done
1	Sameer Sawant	1st week of February	Implementing 1st module/ functionality (*nltk.py/ By making on this module we were able make the input and output of words in array, by tokenizining and stemming and coverting it into numbers)*
		2nd week of February	Testing 1st module (*train.py/ By using intents.json we were able process the data and also train the model and get loss and accuracy*)
2	Raj Shisode	3rd week of February	Implementing 2nd module/ functionality (*model.py/ We build a neural network to predict and train and test model and get data preprocessed accordingly*)
3	Siddhart Chhoriya	By the end of march month	Implementing 3rd module/ functionality (*chat.py/Making the gui and connecting the different module to each other, and displaying the result as working chatbot*)

10. GANTT CHART



Gantt chart help teams to plan work around the deadlines and properly allocate resources. Projects planners also use Gantt charts to maintain a bird's eye view of projects. They depict, among other things, the relationship between the start and end dates of tasks, milestones, and dependent tasks. Modern Gantt chart programs such as Jira Software with Roadmaps and Advanced Roadmaps synthesize information and illustrate how choices impact deadlines.

Chapter 11:

Conclusion

An E-mail spam detector is a great tool for saving time and identifying the original and authentic email. It is developed to provide an accurate response of given input in an easy manner. This project help us to understand training and testing of model along with response and pattern of data processing for email spam detection. While enabling email filtering can prevent phishing and scam emails from victimizing you, sometimes, a few malicious ones can make their way through. Email filters are not enough to protect you from being victimized by a scam. There are a few common signs that an email may be a scam or phishing attempt. The first step is knowing the difference between a Ham email and a spam email.

Working on this project also helps us to provide brief exposure in the domain of Machine Learning and how it will shape our lives in future. Machine learning is now becoming the main part of every technology based enterprise, for getting the user to do automated tasks. Also it helps to give user more information of their product or services in detailed manner.

References

[1] https://www.soscanhelp.com/blog/what-is-email-spam-filtering-and-how-doesit-work.

[2] https://www.mailchannels.com/what-is-spam-filtering/.

[3] Documentation of Pytorch and Nltk

[4] https://www.researchgate.net/publication/342113653_Email_based_Spam_Detection

[5] https://towardsdatascience.com/email-spam-detection-1-2-b0e06a5c0472

[6] https://jpinfotech.org/email-spam-detection-using-machine-learning-algorithms/