Department of Information Technology
NBA Accredited

A.P. Shah Institute of Technology

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

UNIVERSITY OF MUMBAI

Academic Year 2020-2021

A Project Report on

Email Spam Detection

Submitted in partial fulfillment of the degree of
Bachelor of Engineering(Sem-8)
in
INFORMATION TECHNOLOGY

By
Raj Shisode(19104070)
Sameer Sawant(19104054)
Siddhart Chhoriya(20204005)

Under the Guidance of
Apeksha Mohite

1.Project Conception and Initiation

Introduction:

- In the era of information technology, information sharing has become very easy and fast. Many platforms are available for users to share information anywhere across the world.
- Among all information sharing mediums, email is the simplest, cheapest, and the most rapid method of information sharing worldwide. But, due to their simplicity, emails are vulnerable to different kinds of attacks, and the most common and dangerous one is spam.
- No one wants to receive emails not related to their interest because they waste receivers' time and resources.

1.1 Objectives:

- The objective of this Project is to try several classification methods and combination of all methods in order to ultimately choose the one that gives the best results
- To predict whether an email is spam or not using content.
- To understand the working of Machine learning and Artificial Intelligence with respect to this project.

Literature Review:

Title	Author	Findings
Study of Spam E mails	Prince Raj, 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
A systematic literature review on spam content detection and classification	Abbas.M and Ibrahim.M (2014)	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

1.3 Problem Definition

- Unwanted ,junk E mails are piled up in the email box
- Critical Emails are missed or delayed.
- Criminal Minded People are using them for illegal and unethical conducts, phising and fraud.Sending malicious link through spam emails which can harm our system and can also seek in to you system.

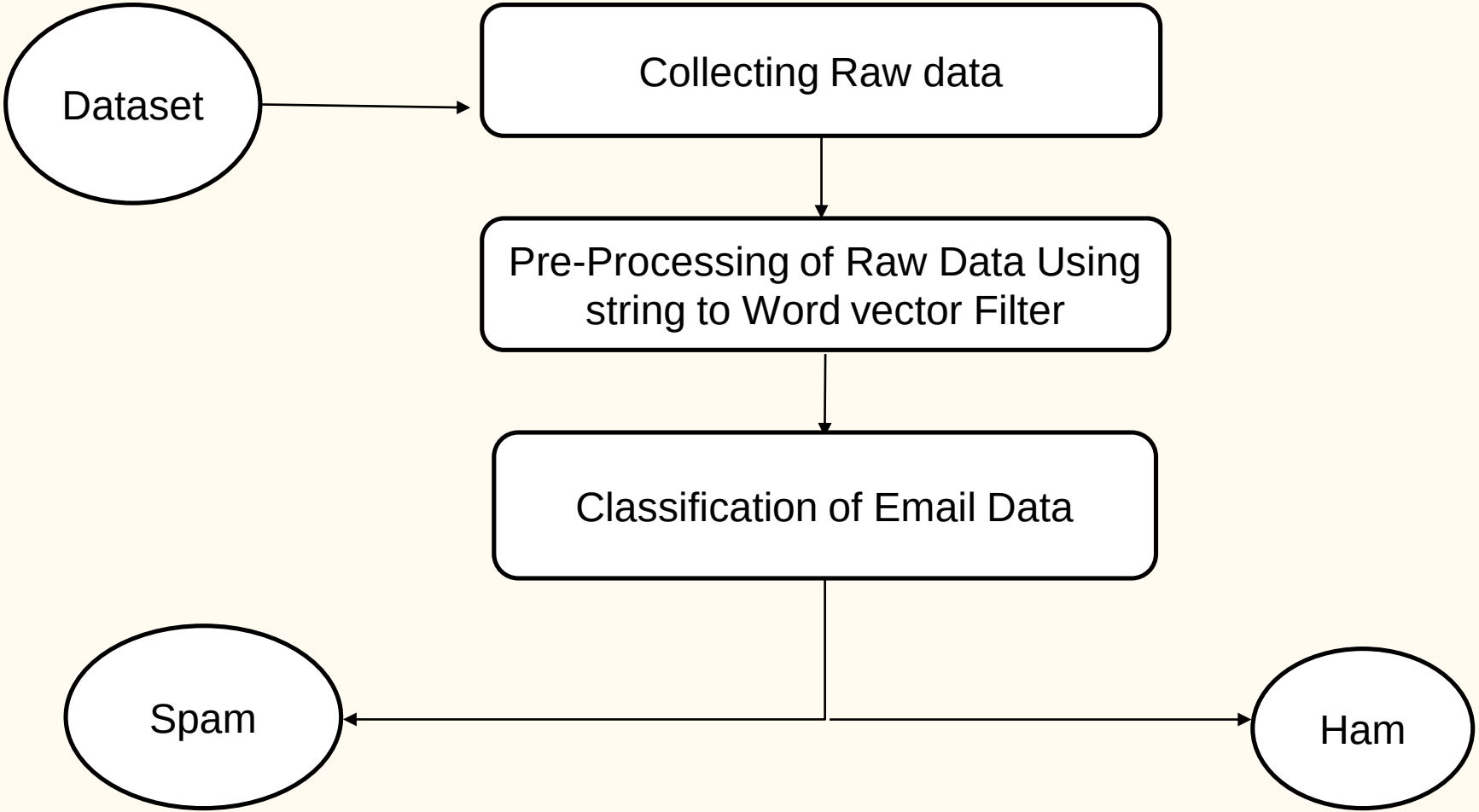
1.4 Scope

- Can be helpful to reject mails that are not important
- Can be used to save time as it helps us to detect whether an email is a spam or a ham.
- Can be used to predict if an email/emails is send by a unethical hacker or not.

1.5 Technology stack

- Python, Streamlit and heroku for hosting website.
- Numpy, Pandas, NLTK(Natural Language Toolkit).
- Naïve bayes, SVM, ETC(Random Forest), etc Algorithms/Modules.

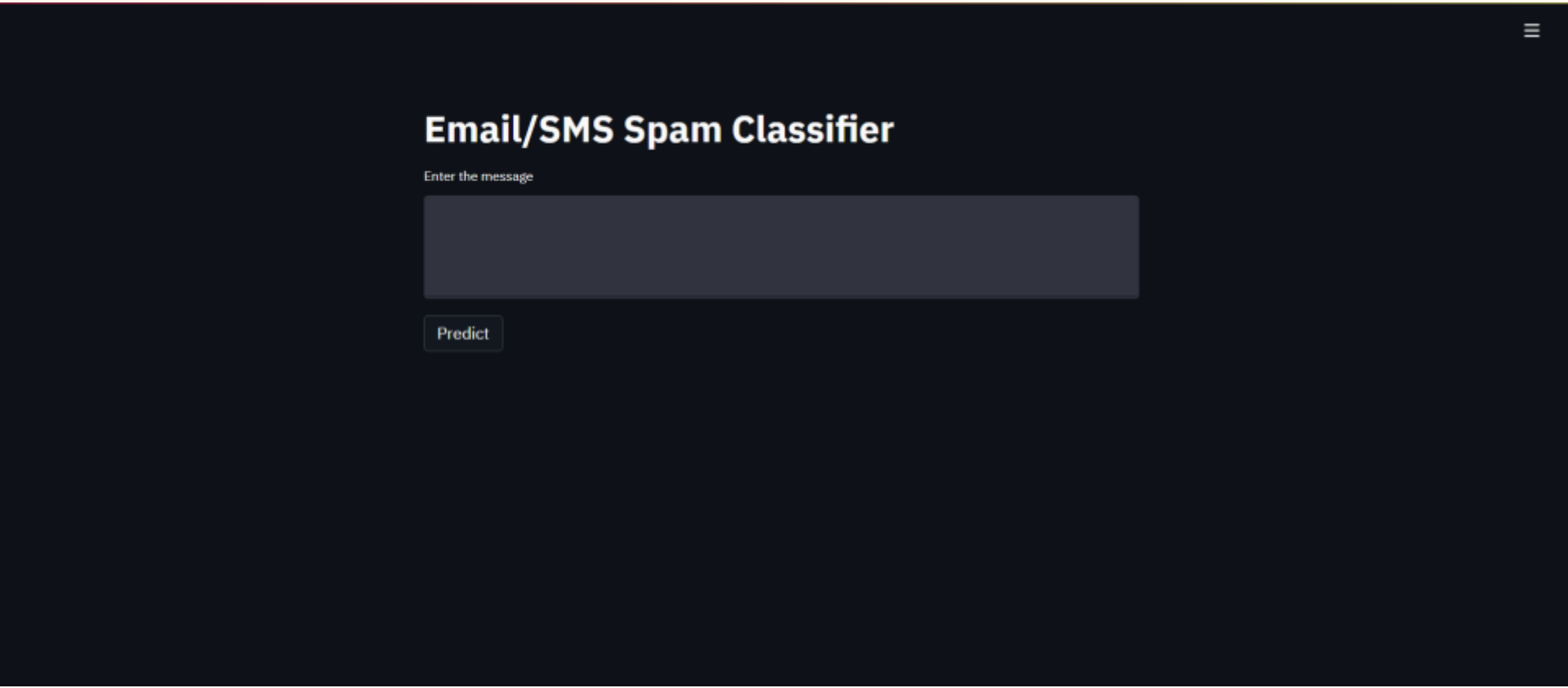
2. Project Design

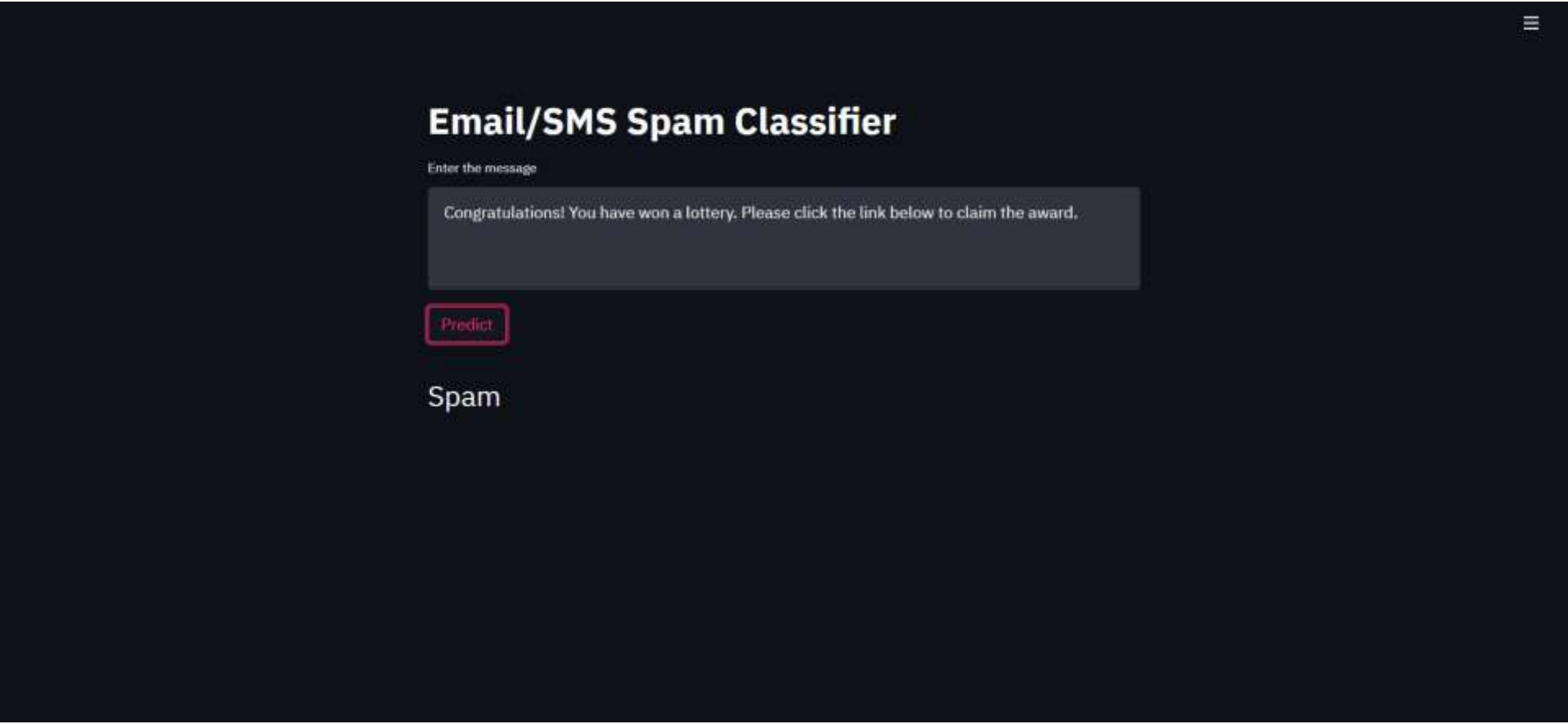

Dataset
Collecting Raw data
Pre-Processing of Raw Data Using string to Word vector Filter
Classification of Email Data
Spam
Ham

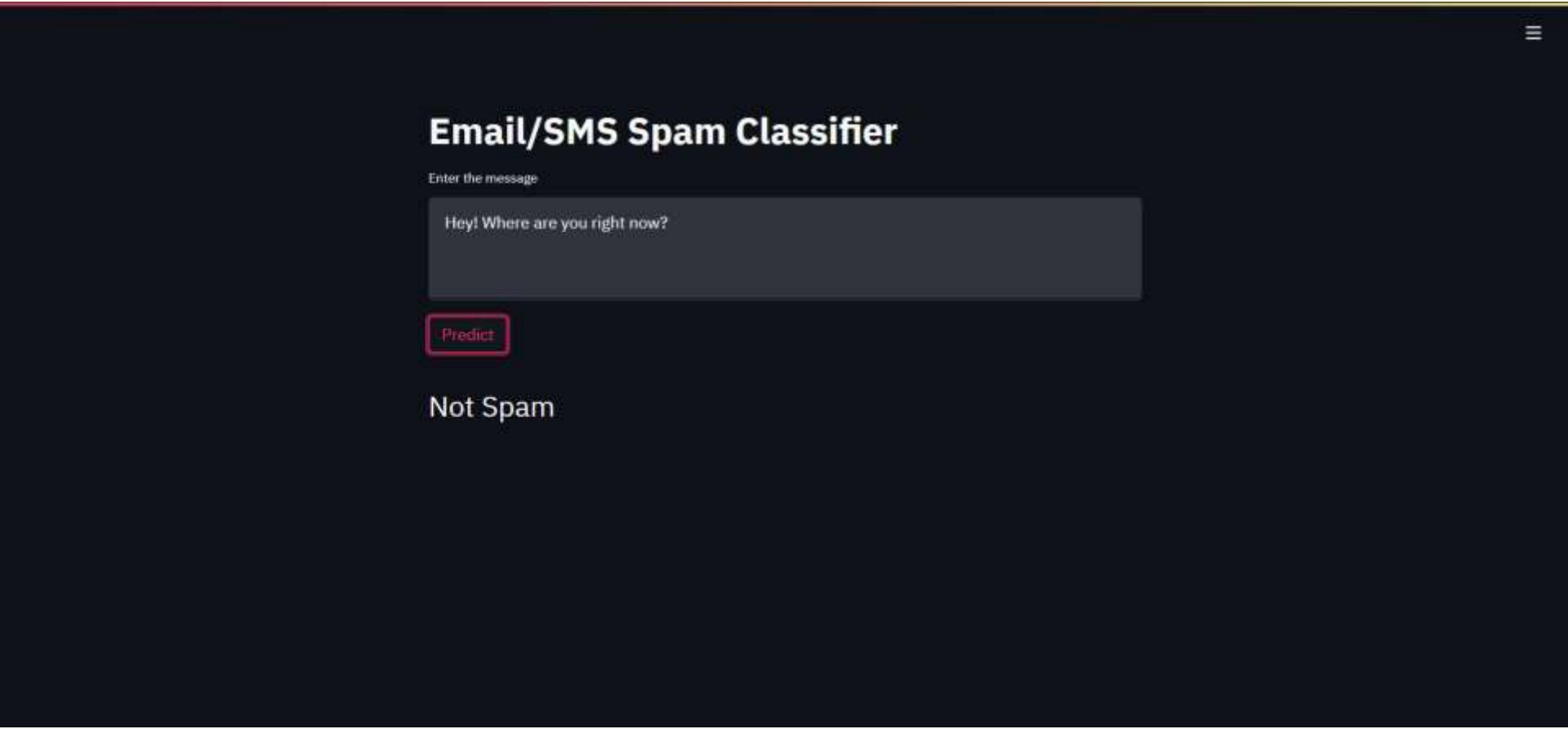
2.1 Proposed System

- An Interface will pe provided with an area to fill up the written content.
- User has to copy the email content and paste it in the textarea .
- We can select the predict button ,to see whether the email infact is a spam or a ham.

3. Implementation


Email/SMS Spam Classifier
Enter the message
Predict


Email/SMS Spam Classifier
Enter the message
Congratulations! You have won a lottery. Please click the link below to claim the award.
Predict
Spam


Email/SMS Spam Classifier
Enter the message
Hey! Where are you right now?
Predict
Not Spam

5. Result

- Our model has been trained using multiple classifiers to check and compare the results for greater accuracy and precision.
- The dataset is obtained from “Kaggle” website for training. The name of the dataset used is “spam.csv”
- Intitially we compared all the algorithms in jupyter notebook, and then we implemented the best algorithm for spam email detection with highest accuracy and precision to check whether the data is “spam” or “ham”.

6. Conclusion and Future Scope

- An E-mail spam detector is a great tool for saving time and identifying the original and authentic email it’s developed to provide an accurate response of input in an easy manner.

- This project help us to understand training and testing of model along with response and pattern data processing for Spam email detector.

- While enabling email filtering can prevent phishing and scam emails from victimizing you, sometimes, a few malicious mails ones can make their way through. Email filters are not enough to protect you from being victimized by a scam.

References

- 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.
- 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
- 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.
- Harisinghaney, Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty, and Information Technology (ICROIT), 2014 International Conference on, pp.153-155. IEEE, 2014

Thank You