e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

RESULT OF EMAL SPAM FILTERING SYSTEM

Prof. Pramod T. Talole[*1], Vaishnavi G. Sahane[*2], Rakhi S. Zade[*3], Poonam D. Kamble[*4], Rohan D. Shingne[*5]

[*1,2,3,4,5]Anuradha Engineering College Chikhli, Buldhana, India.

ABSTRACT

Email spam remains a persistent issue despite various attempts to mitigate its impact. In this paper, we propose an Email Spam Filtering System leveraging machine learning techniques. The system is designed to classify incoming emails as spam or non-spam (ham) using a trained model. We present the methodology, module description, and future scope of our system, highlighting its potential to enhance email security and productivity. Additionally, we provide insights into the implementation details and discuss the effectiveness of our approach in combating email spam. In this project, machine learning techniques are used to detect the spam message of a mail. Machine learning is where computers can learn to do something without the need to explicitly program them for the task. It uses data and produce a program to perform a task such as classification. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in machine learning studio.

Keywords: Email, Spam, Machine Learning, Pandas, Streamlit, Numpy, Naïve Bias.

I. INTRODUCTION

Email spam poses significant challenges to individuals and organizations, ranging from privacy concerns to resource wastage. Traditional rule-based spam filters often struggle to keep pace with evolving spamming techniques. Therefore, we introduce an advanced Email Spam Filtering System based on machine learning algorithms. By automatically categorizing incoming emails, our system aims to streamline inbox management and improve overall email security. Today, Spam has become a major problem in communication over internet. It has been accounted that around 55% of all emails are reported as spam and the number has been growing steadily. Spam which is also known as unsolicited bulk email has led to the increasing use of email as email provides the perfect ways to send the unwanted advertisement or junk newsgroup posting at no cost for the sender. This chance has been extensively exploited by irresponsible organizations and resulting to clutter the mail boxes of millions of people all around the world. Spam has been a major concern given the offensive content of messages, spam is a waste of time. End user is at risk of deleting legitimate mail by mistake. Moreover, spam also impacted the economical which led some countries to adopt legislation. Text classification is used to determine the path of incoming mail/message either into inbox or straight to spam folder. It is the process of assigning categories to text according to its content.

1.1 PROJECT AIMS AND OBJECTIVES

The project aims and objectives that will be achieved after completion of this project are discussed in this subchapter. The aims and objectives are as follows:

- Accuracy: Our aim is to achieve a minimum accuracy rate of 95% in classifying emails as spam or non-spam, validated through rigorous testing against benchmark datasets and real-world email traffic.
- False Positive Rate: We're committed to keeping the false positive rate under 1%, ensuring legitimate emails aren't erroneously flagged as spam.
- False Negative Rate: Minimizing the false negative rate to less than 5% guarantees spam emails are accurately identified and filtered out to protect users from potential security threats.
- Response Time: Maintaining a response time of under 1 second for processing and filtering incoming emails ensures near real-time detection and filtering of spam.
- User Satisfaction: Through user surveys and feedback sessions, we're targeting a satisfaction rate of at least 90% to ensure the spam filtering system meets user expectations.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

II. METHODOLOGY

2.1. SYSTEM ANALYSIS

This chapter will explain the specific details on the methodology being used to develop this project. Methodology is an important role as a guide for this project to make sure it is in the right path and working as well as plan. There is different type of methodology used in order to do spam detection and filtering.

So, it is important to choose the right and suitable methodology thus it is necessary to understand the application functionality itself.

2.1.1 SOFTWARE AND HARDWARE REQUIREMENTS

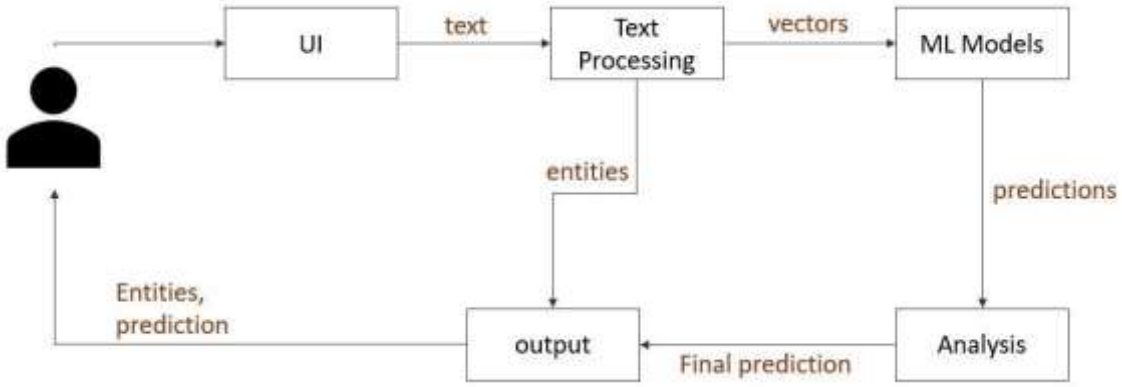
Software Requirements

- OS – Windows 7 and above
- Code Editor – Pycharm, VS Code,
- Built in IDE Anaconda environment with packages nltk, numpy, pandas, sklearn, tkinter, nltk data.
- Supported browser such as chrome, firefox, opera etc.

Hardware Requirements

- PC/Laptop Ram – 8 GB
- Storage – 100-200 Mb

Architecture



III. MODULE DESCRIPTION

The Application consists of three modules.

i. UI

ii. Machine Learning

iii. Data Processing

I. UI Module

a. This Module contains all the functions related to UI(user interface).

b. The user interface of this application is designed using Streamlit library from python based packages.

c. The user inputs are acquired using the functions of this library and forwarded to data processing module for processing and conversion.

d. Finally the output from ML module is sent to this module and from this module to user in visual form.

II. Machine Learning Module

a. This module is the main module of all three modules.

b. This modules performs everything related to machine learning and results analysis.

c. Some main functions of this module are

i. Training machine learning models.

ii. Testing the model

iii. Determining the respective parameter values for each model.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

iv. iv. Key-word extraction.

v. Final output calculation

d. The output from this module is forwarded to UI for providing visual response to user

III. Data Processing Module

a. The raw data undergoes several modifications in this module for further process.

b. Some of the main functions of this module includes

i. Data cleaning

ii. Data mergingof datasets

iii. Text Processing using NLP

iv. Conversion of text data into numerical data(feature vectors).

v. Splitting of data.

c. All the data processing is done using Pandas and NumPy libraries.

d. Text processing and text conversion is done using NLTK and scikit-learn libraries

System Testing

The aim of the system testing process was to determine all defects in our project. The program was subjected to a set of test inputs and various observations were made and based on these observations it will be decided whether the program behaves as expected or not. Our Project went through two levels of testing.

Unit testing.

Integration testing.

Unit testing

Unit testing is undertaken when a module has been created and succesfully reviewed in order to test a single module we need to provide a complete environment ie besides the module we would require.

- The procedures belonging to other modules that the module under test calls.
- Non-local data structures that module accesses.
- A procedure to call the functions of the module under test with appropriate parameters.

Integration Testing

In this type of testing we test various integration of the project module by providing the input the primary objective is to test the module interfaces in order to ensure that no errors are occurring when one module invokes the other module.

IV. FUTURE SCOPE

There are numerous applications to machine learning and natural language processing and when combined they can solve some of the most troubling problems concerned with texts. This application can be scaled to intake text in bulk so that classification can be done more affectively in some public sites. Other contexts such as negative, phishing, malicious, etc,. can be used to train the model to filter things such as public comments in various social sites. This application can be converted to online type of machine learning system and can be easily updated with latest trends of spam and other mails so that the system can adapt to new types of spam emails and texts. While the current iteration of the Email Spam Filtering System showcases promising functionality, there are several avenues for future 60 enhancement and development. One potential area for improvement lies in refining the machine learning model to achieve even higher accuracy in email classification. This could involve exploring advanced algorithms or incorporating additional features to better capture the nuances of spam emails. Furthermore, the scalability of the system could be enhanced to accommodate large volumes of incoming emails without compromising performance. Implementing parallel processing techniques or deploying the system on cloud infrastructure could facilitate this scalability, catering to the needs of organizations with diverse email traffic. Additionally, the integration of natural language processing (NLP) techniques could enable the system to analyse email content more comprehensively, considering semantic meaning and context. This could further improve the system's ability to accurately

identify spam emails, especially those employing sophisticated obfuscation techniques.

V. CONCLUSION

In conclusion, our Email Spam Filtering System presents a promising approach to tackle the pervasive issue of email spam. By leveraging machine learning techniques and a user-friendly interface, our system provides an efficient means of identifying and mitigating spam emails. With further enhancements and integration possibilities, our system holds the potential to significantly improve email security and streamline communication workflows.

VI. REFERENCES

[1] S. H. a. M. A. T. Toma, "An Analysis of Supervised Machine Learning Algorithms for Spam Email Detection," in International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021.

[2] S. Nandhini and J. Marseline K.S., "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection," in International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020. 62

[3] A. L. a. S. S. S. Gadde, "SMS Spam Detection using Machine Learning and Deep Learning Techniques," in 7th International Conference on Advanced Computing and Communication Systems (ICACCS), 2021, 2021.

[4] V. B. a. B. K. P. Sethi, "SMS spam detection and comparison of various machine learning algorithms," in International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017.

[5] G. D. a. A. R. P. Navaney, "SMS Spam Filtering Using Supervised Machine Learning Algorithms," in 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018.

[6] S. O. Olatunji, "Extreme Learning Machines and Support Vector Machines models for email spam detection," in IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017.

[7] S. S. a. N. N. Kumar, "Email Spam Detection Using Machine Learning Algorithms," in Second International Conference on Inventive Research in Computing Applications (CIRCA), 2020.

[8] R. Madan, "medium.com," [Online]. Available: https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanatio n-fortext-classification-in-nlp-with-code-8ca3912e58c3.

[9] N. D. J. a. M. M. A. M. M. RAZA, "A Comprehensive Review on Email Spam Classification using Machine Learning Algorithms," in International Conference on Information Networking (ICOIN), 2021, 2021.

[10] A. B. S. A. a. P. M. M. Gupta, "A Comparative Study of Spam SMS Detection Using Machine Learning Classifiers," in Eleventh International Conference on Contemporary Computing (IC3), 2018.

[11] M. M. J. Fattahi, "SpaML: a Bimodal Ensemble Learning Spam Detector based on NLP Techniques," in IEEE 5th International Conference on Cryptography, Security and Privacy (CSP), 2021, 2021.

[12] Harika, "Analytics Vidhya," [Online]. Available: https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/.

[13] İ. A. D. a. M. D. H. Karamollaoglu, "Detection of Spam E-mails with Machine Learning Methods," in Innovations in Intelligent Systems and Applications Conference (ASYU), 2018.

[14] M. N. U. a. R. K. H. F. Hossain, "Analysis of Optimized Machine Learning and Deep Learning Techniques for Spam Detection," in IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS), 2021.

[15] H. Deng, "Towards Data Science," [Online]. https://towardsdatascience.com/random-forest-3a55c3aca46d