**Spam Email Classification using NLP and Machine Learning**

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning

with

TechSaksham – A joint CSR initiative of Microsoft & SAP

by

**ARYAN DEVANAND SHINDE, aryanshinde2022.comp@mmcoe.edu.in**

Under the Guidance of

**Abdul Aziz Md**

**Master Trainer, Edunet Foundation**

**ACKNOWLEDGEMENT**

I wish to thank Mr. Abdul Aziz for the immense guidance, support, and encouragement he has provided during my one-month internship. He has been a great source of learning for me in understanding many of the key concepts and applications, especially in AI technologies. His constructive feedback and insightful advice have been crucial in my personal and professional growth during this period.

I am deeply grateful to the joint CSR initiative of Microsoft and SAP for giving me the opportunity to be part of such a powerful program. This initiative, focused on AI technologies, has given me access to the latest resources and a platform to work on innovative solutions. It not only broadened my technical expertise but also deepened my appreciation for the role of technology in solving real-world challenges.

I would like to acknowledge the invaluable support and encouragement from my fellow peers, who were ever willing to share knowledge and collaborate effectively during this internship. Finally, I extend the deepest gratitude to my family and friends for their constant motivation, understanding, and belief in my abilities, which have been a great source of strength throughout this journey.

#### **ABSTRACT**

The growth of unwanted and malicious emails has become a serious issue that requires efficient and automated methods for spam detection. This project is concerned with developing a machine learning-based solution to classify emails as spam or not spam using Natural Language Processing (NLP) techniques.

The main goal for the project is to develop an effective and accurate spam detection system by processing and classifying its content. The problem is solved by using a supervised dataset, in which labeled data of emails (spam as well as non-spam) is used as training data for the model.  
  
This involves pre-processing the text from emails with methods such as tokenization, removal of stop-words, and vectorization methods, such as TF-IDF. These convert raw text from emails into structured input for machine learning algorithms. For classification, Random Forest is used as an effective classifier in handling high-dimensional data, which also overcomes the problem of overfitting due to ensemble learning. Naïve Bayes technique is used to built the model.  
  
The performance of the model is measured in terms of accuracy, precision, recall, and F1-score. The results show high accuracy with reliable performance in the identification of spam emails against non-spam emails, highlighting the effectiveness of the Random Forest algorithm for this kind of classification task.  
  
The project thus concludes by applying the techniques of machine learning and NLP for the development of the challenge of detecting spam email. It can thus be enhanced by adding new features like real-time classification, more sophisticated NLP methods to increase robustness and scalability.

**TABLE OF CONTENTS**

Abstract 0

**Chapter 1.**  **Introduction**

1.1 Problem Statement 7

1.2 Motivation 8

1.3 Objectives 9

1.4. Scope of the Project 9

**Chapter 2.**  **Literature Survey 10**

**Chapter 3.**  **Proposed Methodology 11**

**Chapter 4.**  **Implementation and Results 13**

**Chapter 5. Discussion and Conclusion 16**

**References** 17

**LIST OF FIGURES**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Fig sr** | **NAME** | **Page No.** |
|  | 1.1 System Architecture | **11** |
|  | 1.2 Not spam image | **13** |
|  | 1.3 Spam image | **13** |
|  | 1.4 Analysis | **14** |
|  | 1.5 Analysis |  |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table. No.** | **Table Caption** | **Page No.** |
| **1.1** | Workdone on concept | **10** |
| **1.2** | Hardware Requirements | **12** |
| **1.3** | Software Requirements | **12** |

**CHAPTER 1**

**Introduction**

* 1. **Problem Statement:**

Spam email classification is the process of identifying and separating unwanted or unsolicited emails, referred to as spam, from legitimate ones, known as ham. This is done using Natural Language Processing (NLP) and machine learning (ML) techniques to analyze email content, sender details, and metadata in order to distinguish between the two.

**Significance of the Problem**

**Impact on Productivity:**

Spam emails fill up inboxes requiring users to spend a lot of time filtering and deleting those unnecessary messages, thus eroding productivity.

**Security Breaches:**

Most spam contains malicious content, phishing link, or malware that can undermine personal or organizational data by resulting in financial and reputation losses.

**Data Overload:**

The volume of spams with billions of these emails daily adds more pressures on email servers and networks, hence increasing operational expenses.

It helps improve the user experience by only delivering relevant and legitimate emails to the inbox.

**Legal Compliance:**

Most countries have strict regulations against spam emails, such as the GDPR and CAN-SPAM Act. Accurate classification will ensure compliance and prevent fines.

**Evolving Threats:**

Spammers continuously develop more sophisticated techniques to evade filters. Advanced NLP and ML solutions are necessary to keep up with these threats.

* 1. **Motivation:**

This project has been selected because the frequency of spam emails is growing with their sophistication. This task of spam classification through state-of-the-art techniques like NLP and ML poses a challenge to make cybersecurity robust. Moreover, the project provides an avenue for applying cutting-edge concepts of AI/ML to a real-world problem that touches upon people, businesses, and governments around the world.

**Applications**

Email Service Providers:

Integrate spam detection with email providers, such as Gmail and Outlook, to improve accuracy for preventing malicious emails.

**E-commerce Solutions:**

Reduces fraudulent or spam promotional emails

**Government Institutions:**

Spam campaign against citizens and the institutions of the nation would be detected to prevent hacking of national security systems.

**Research and Development:**

This is a basic research that would help in the development of smarter AI systems that can handle multi-lingual and context-specific spam filtering.

Impact

**Cost Savings:**

Reduces server and storage costs in handling and processing spam emails.

**Trust and Reliability:**

Increases trust in email as a communication medium, especially in e-commerce and financial transactions.

* 1. **Objective:**

The major goal of this project is to design a spam classifier using machine learning technique. It would automatically classify the input email as either spam or non-spam (ham). This system would check all the content, meta information, and other features of a particular incoming email to identify pattern characteristics and features associated with emails identified as spam. Therefore, the objective in designing an email filtering system includes improving the efficiency of rejecting unwanted emails and avoiding spam messages.

Data Collection: Gather and preprocess a labeled dataset of emails.

Feature Extraction: Extract relevant features from the content of emails (words, phrases, sender information, etc.).

Train machine learning models such as Naive Bayes, SVM, Random Forest on the dataset.

Evaluation: Model performance should be evaluated in terms of accuracy, precision, recall, and F1-score.

* 1. **Scope of the Project:**

Data Processing: Gathering and preprocessing a labeled dataset of emails (spam and non-spam).

Model Development: Training models (e.g., Naive Bayes, SVM) on the dataset.

Spam Detection: Classifying emails as spam or non-spam.

Evaluation: Evaluating the model's performance using metrics like accuracy, precision, recall, and F1-score.

**Limitations:**

Dataset Limitation: model performance is based on data quality and size.

Feature selection: Some email features, as images or attachments, not included.

Complex Spam Tactics: It may be hard for the system to detect advanced spam methods.

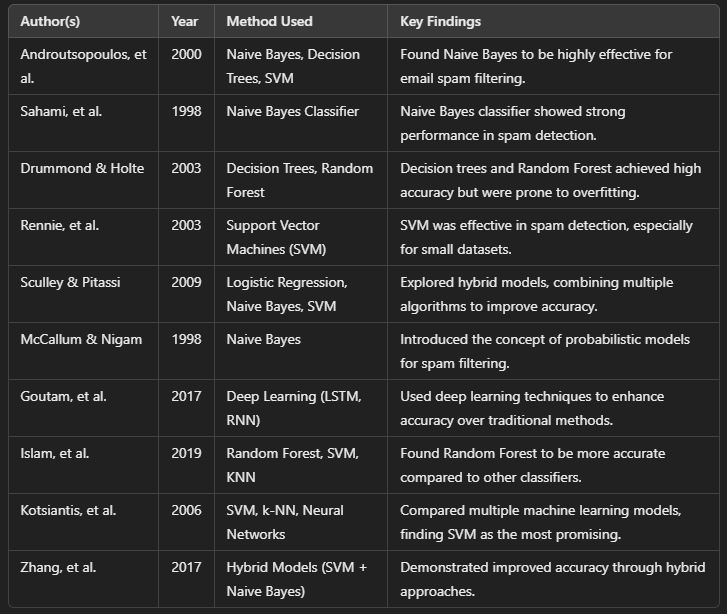
In Real-Time Processing: this system is not designed in real-time spam detection in live Email systems.

Language and context: The model will have the problem of context-based spam as well as non-English emails.

**CHAPTER 2**

**Literature Survey**

**Previous Workdone on this concept using various Methods:**

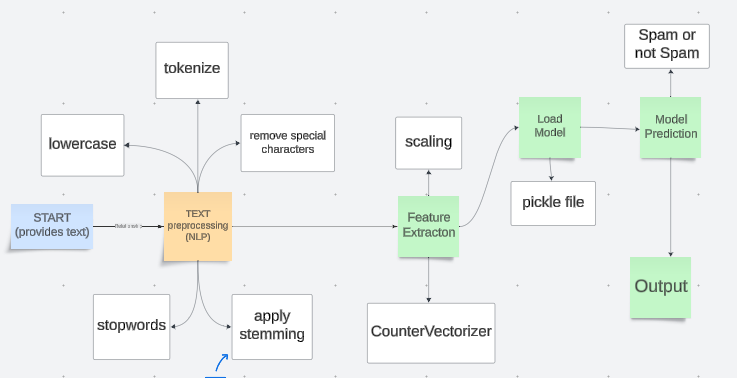
****

*(table 1.1)*

**CHAPTER 3**

**Proposed Methodology**

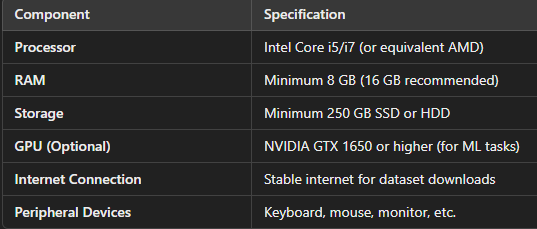
* 1. **System Design**

****

*(fig 1.1)*

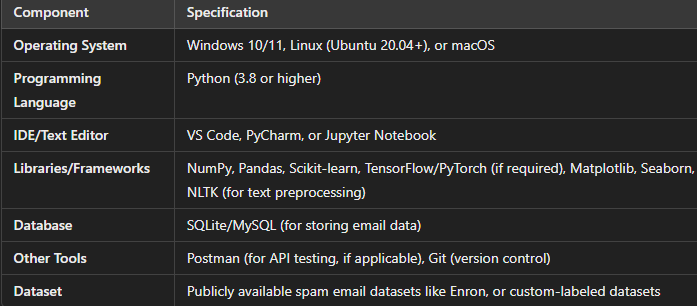
The above fig is of System Architecture for Spam Email detection which is done using NLP.

* 1. **Requirement Specification**
     1. **Hardware Requirements:**



*(table 1.2)*

**3.2.2 Software Requirements:**

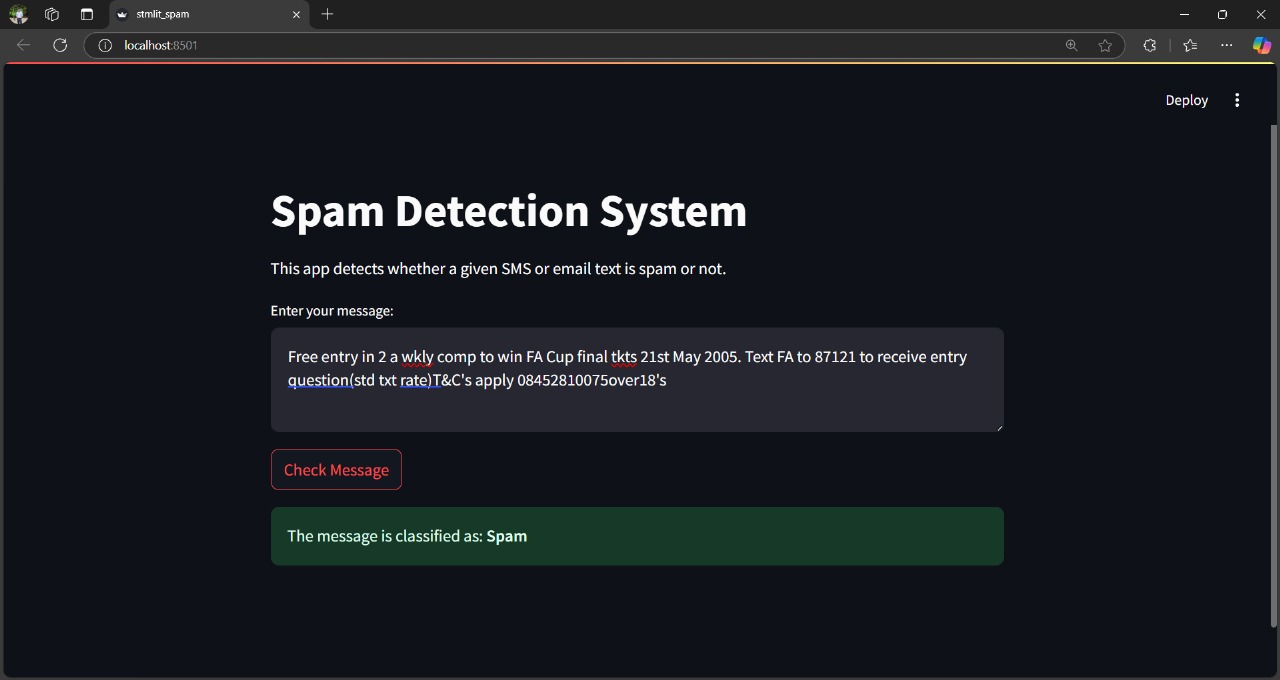
****

*(table 1.3)*

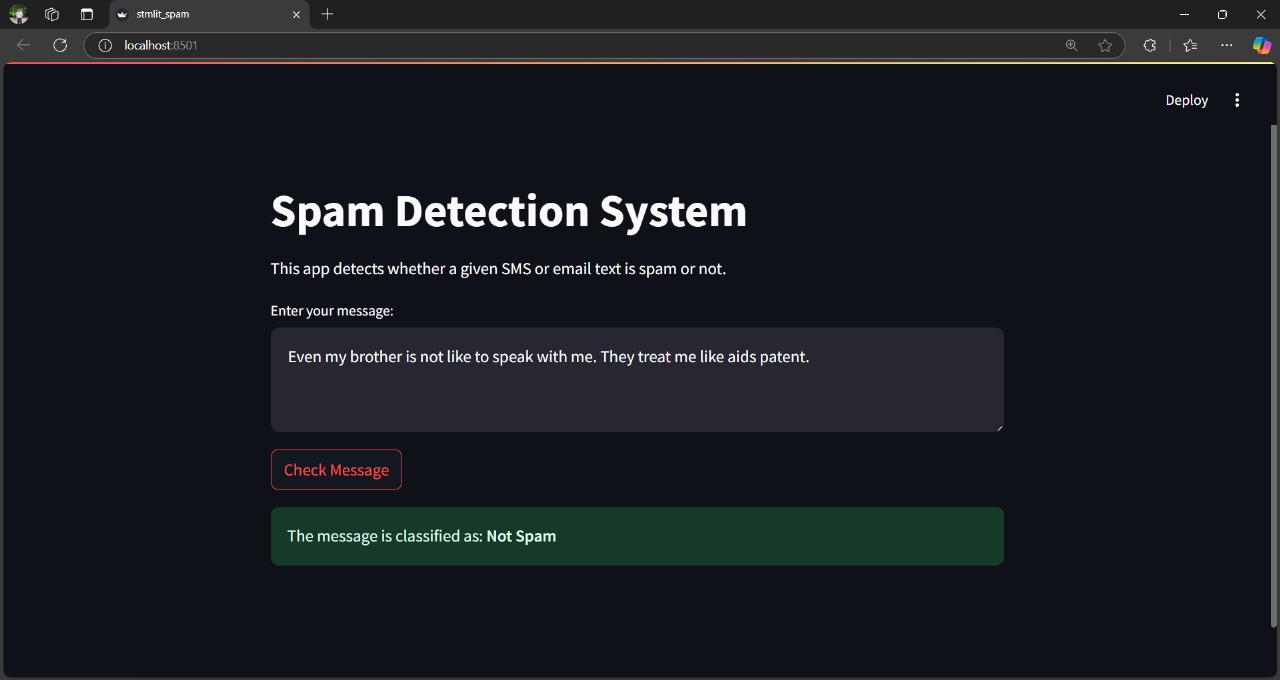
**CHAPTER 4**

**Implementation and Result**

* 1. **Snap Shots of Result**

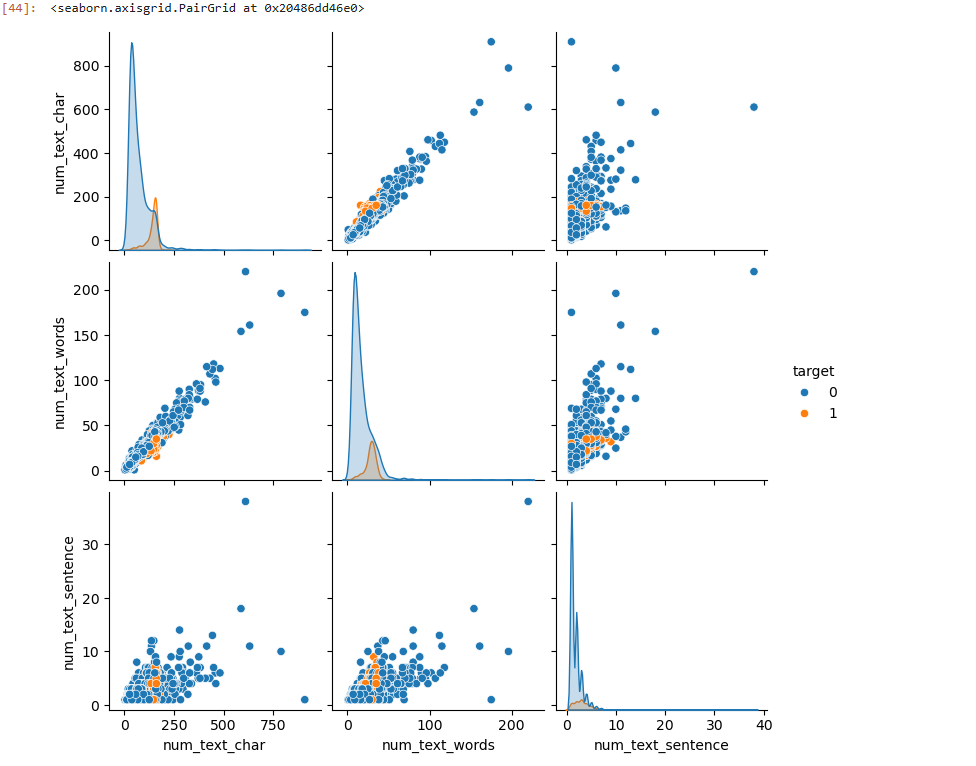


*(fig 1.2)* shows the output of spam



*(fig 1.3)* shows the output of not spam

**ANALYSIS:**



*(fig 1.4) analysis*



*(fig 1.5) analysis*

* 1. **GitHub Link for Code:**

[**Aryan-Devanand-Shinde/Spam-Email-Classification-using-NLP-and-Machine-Learning-: Spam Email Classification using NLP and Machine Learning is an project done in internship at edunet Foundation**](https://github.com/Aryan-Devanand-Shinde/Spam-Email-Classification-using-NLP-and-Machine-Learning-)

**CHAPTER 5**

**Discussion and Conclusion**

* 1. **Future Work:**

Future work for the email spam detection project can focus on improving model performance through advanced techniques like deep learning, ensemble methods, and multi-language support. Expanding and updating datasets with real-time data collection will enhance adaptability to evolving spam patterns. Incorporating adversarial training and malware analysis can strengthen security, while explainable AI methods can improve transparency and trust. Real-time deployment, scalability via cloud integration, and integration with systems like NIDS can further extend functionality. Finally, adding robust evaluation metrics and visual analytics will refine accuracy and usability in practical applications.

* 1. **Conclusion:**

In conclusion, improving the email spam detection model involves enhancing its accuracy, adaptability, and security through advanced techniques, robust datasets, and real-time deployment. Integrating explainability, scalability, and additional features like malware analysis will make the system more effective and reliable for real-world applications.

**REFERENCES**

1. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.
2. M. P. Prabhakaran, B. Abhijit, A. H. Vinayak, and P. S. Divya, "A Comprehensive Survey for Intelligent Spam Email Detection," *IEEE Transactions on Big Data and Informatics*, Vol. 35, No. 2, pp. 110-125, 2019
3. P. T. Suresh and K. Anand, "Detection of Email Spam using Machine Learning Algorithms: A Comparative Study," *IEEE Conference on Emerging Trends in Computer and Information Sciences*, Vol. 12, pp. 210-215, 2024
4. A. K. Sharma and R. Patel, "Machine Learning Approaches for Automatic Email Spam Detection," *IEEE Conference on Recent Trends in Artificial Intelligence and Machine Learning*, Vol. 28, No. 4, pp. 190-195, 2024
5. S. Gupta, M. Verma, and K. Das, "Email Spam Detection using Deep Learning Approach," *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 45, No. 6, pp. 320-330, 2024