**Mini Project**

**FOR**

**Spam Message Filtering**

**(Semester – V)**

**(Computer Engineering Department)**

**Prepare By:- Enrollment No:-**

**Tirth Patel 21002170310038**

**Academic Year(2023-24)**

**Mr. Krunal Panchal.**

**(Faculty Guide)**

****

**L. J. Institute of Engineering & Technology, Ahmadabad (LJU)**



**L. J. INSTITUTE OF ENGINEERING AND TECHNOLOGY**

**TY-Department 2023 – 24**

**CERTIFICATE**

Date: 16-Jan-2024

This is to certify to **Tirth Patel** of B.E Semester **5thI.T.** Class, Enrollment No. **21002170310038** has satisfactorily completed his Mini Project work of the subject **Spam Message Filtering** during the academic year **2023-24** and submitted on**16-jan-2024**

**Head of Department Guided By**

**Prof. Shruti Raval Prof. Krunal Panchal**

**Computer Department Computer Department**

**LJU, Ahmedabad LJU, Ahmedabad**

Guide

Mr. Bhavesh Vaghela

TY – Department, L.J.I.E.T.

INDEX

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Title** | **Page No.** |
|  |  |  |
| 1. | Acknowledgement |  |
| 2. | Abstract |  |
|  |  |  |
| 3. | [Introduction](#OverallDescription) |  |
|  | 3.1 Background |  |
|  | 3.2 Motivation |  |
|  | 3.3 Objective |  |
|  | 3.4 Challenges |  |
|  |  |  |
| 4. | Literature Review |  |
| 5. | Implementation model |  |
|  |  |  |
| 6. | Methodology |  |
|  | 6.1 Class Diagram |  |
|  | 6.2 Use Case Diagram |  |
|  | 6.3 Activity Diagram |  |
|  |  |  |
| 7. | Experiments and results |  |
|  | 7.1 Experiment of machine learning algorithm |  |
|  |  |  |
| 8. | Conclusion |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**CHAPTER – 1**

**INTRODUCTION**

# 1. ACKNOWLEGEMENT

We are heartily thankful to all faculty members of the department of Computer Engineering from L.J University, Ahmedabad for making my project. It is my pleasure to take this opportunity to thank all people who helped me directly or indirectly to prefer this project would have been impossible without their guidance. They all encouraged and trusted in our ideas. They were always available for us to give guidance about the project. The disruption about the project and the great advice given by them helped to make this project complete. We are thankful to them pristine and enlightening guidance given to as throughout the semester.

We are especially thankful to our internal guide **Prof. Krunal Panchal** , for their Encouragement, guidance, understanding and lots of support and trust. Without his help this project would not be success. Finally, we thank all persons who directly or indirectly supported us in making this project.

# 2. ABSTRACT

# With the increasing volume of electronic communication, the proliferation of spam emails has become a pervasive issue, posing threats to user privacy and system security. This research presents a novel machine learning (ML) approach to predict and mitigate spam emails effectively. The proposed system employs a robust set of features extracted from email content, headers, and sender information, leveraging state-of-the-art ML algorithms for classification.

# The feature extraction process involves the analysis of text content using natural language processing techniques, examining patterns in sender behavior, and evaluating the relevance of email headers. These features are then used to train and optimize ML models, including but not limited to Support Vector Machines, Random Forests, and Neural Networks. The selection of the most suitable algorithm is based on comparative performance evaluations.

**CHAPTER – 3**

**Introduction**

3.1 Background

* Historical Evolution:

Spam emails have been a persistent issue since the early days of the internet. The term "spam" in the context of electronic communication originated from a famous Monty Python sketch where the word was repetitively used, drawing a parallel to the unsolicited and repetitive nature of unwanted emails. In the mid-1990s, with the increasing popularity of email communication, spam emails became a prevalent and disruptive phenomenon.

* Techniques Employed by Spammers:

Over the years, spammers have continuously evolved their techniques to bypass traditional email filters. These techniques include obfuscating email content, using image-based text to evade text-based filters, employing URL redirects, and employing social engineering tactics to deceive recipients.

**3.2 Motivation**

* User Experience Improvement:

Spam emails can inundate users' inboxes, causing frustration, wasting time, and diminishing the overall user experience. By accurately predicting and filtering out spam, email users can enjoy a cleaner and more efficient communication environment, allowing them to focus on legitimate and relevant messages.

* Privacy and Security Concerns:

Many spam emails are designed to be malicious, carrying threats such as phishing attempts, malware, and scams. Predicting and blocking these harmful messages is essential for safeguarding user privacy, protecting sensitive information, and preventing potential financial or identity-related damages.

**3.3 Objective**

* Real-time Analysis:

Objective: Implement real-time or near-real-time analysis of incoming emails to promptly identify and filter out spam.

Rationale: Timely detection is essential to minimize the potential impact of malicious activities associated with spam emails, such as phishing or malware distribution.

* Minimize False Positives:

Objective: Reduce the occurrence of false positives to avoid filtering out legitimate emails.

Rationale: Minimizing false positives is essential for maintaining trust in the email system and ensuring that important messages are not mistakenly identified as spam.

3.4 Challenges

* False Positives and False Negatives:

Challenge: Balancing false positives (legitimate emails flagged as spam) and false negatives (spam emails not detected) is a delicate challenge.

Implications: Striking the right balance is critical to avoid inconveniencing users by filtering out important emails (false positives) or allowing spam to reach inboxes (false negatives).

**CHAPTER – 4**

**Literature Review**

* **Machine Learning Approaches:**

A significant portion of the literature focuses on the application of machine learning techniques for spam email prediction. Common algorithms include Support Vector Machines (SVM), Naive Bayes, Decision Trees, Random Forests, and Neural Networks.

Feature extraction plays a crucial role, and studies explore various features from email content, headers, and sender information.

* **Deep Learning Approaches:**

With the rise of deep learning, there is a growing interest in leveraging neural networks for spam email prediction. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are applied to capture complex patterns and relationships in email data.

* **Evolution of Spam Tactics:**

The dynamic nature of spam tactics is recognized, and literature discusses the need for adaptive models that can continuously learn and adapt to new spamming techniques. Zero-day attack scenarios, where new and previously unseen spam tactics emerge, are acknowledged as challenges.

* **Real-time Prediction and Scalability**:

There is an emphasis on the importance of real-time prediction, especially in large-scale email systems. Scalability becomes a key consideration, and studies propose solutions to efficiently process and predict spam in high-volume environments

**CHAPTER – 5**

**Implementation of Model**

* **Data Collection:**

Gather a diverse dataset containing labeled examples of spam and non-spam (ham) emails. Ensure the dataset is representative of the types of emails the model will encounter in the real-world scenario.

* **Data Preprocessing:**

Clean and preprocess the email data. This may involve:

Removing irrelevant information (e.g., email addresses, timestamps).

Tokenizing and stemming words to extract features.

Handling missing or incomplete data.

Balancing the dataset if there is a significant class imbalance

* **Feature Extraction:**

Extract relevant features from the preprocessed data. Common features include:

Bag-of-words representations for email content.

Analysis of sender behavior.

Examination of email headers.

Additional features identified through natural language processing (NLP) techniques.

* **Model Selection**:

Choose a suitable machine learning algorithm for spam email prediction. Common choices include:

Support Vector Machines (SVM).

Naive Bayes.

Decision Trees and Random Forests.

Neural Networks (e.g., deep learning models).

* **Training the Model**:

Split the dataset into training and testing sets.

Train the selected model using the training set. This involves providing the model with labeled examples and adjusting its parameters to learn the underlying patterns in the data

* **Monitoring and Maintenance:**

Implement monitoring tools to track the model's performance in real-world scenarios. This involves regularly checking for false positives and false negatives and adjusting the model as needed.

Update the model periodically to adapt to changes in spam tactics and to maintain optimal performance.

**CHAPTER – 6**

**Methodology**

6.1Class Diagram

A diagram with text and words

Description automatically generated

**6.2 Use Case Diagram**

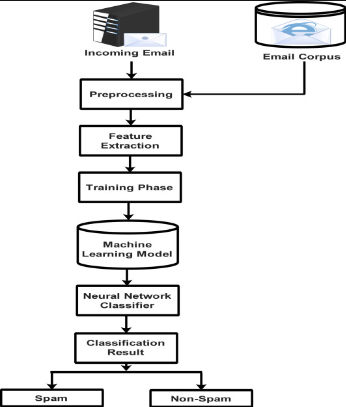
A diagram of a data processing process

Description automatically generated

Use case diagrams are used to gather the requirements of a system including internal and external influences. These requirements are mostly design requirements. Hence, when a system is analyzed to gather its functionalities, use cases are prepared and actors are identified

* This diagram in the there are 5 use cases:
  + Datasets
  + Pre Processing Process
  + Training Datasets
  + Classification
  + Result of Classification

**6.3 Activity Diagram**



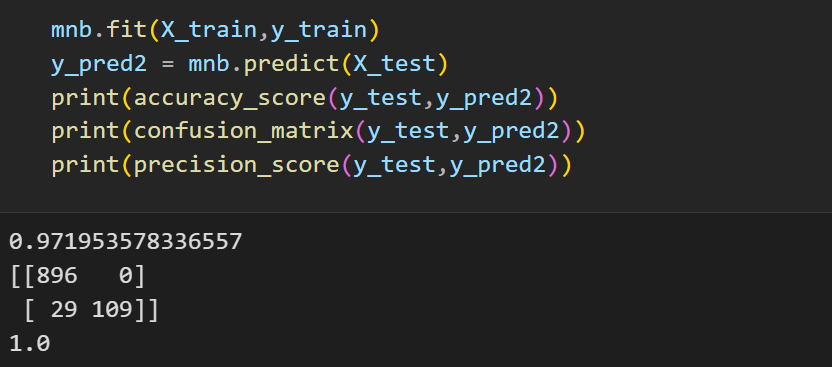
* An activity diagram visually presents a series of actions or flow of control in a system similar to a flowchart or a data flow diagram.
* Activities modeled can be sequential and concurrent. In both cases an activity diagram will have a beginning (an initial state) and an end (a final state).
* Activity Diagram of spam mail filtering is start from initial state or start point. after initial state the flow goes into action state and the first action Is Reading the Incoming Email. then the flow going into other action which is Traning phase.
* When an activity requires a decision prior to moving on to the next activity, add a diamond between the two activities.
* After the classification result email goes either spam mail or ham mail.

**CHAPTER – 7**

**Experiment and result**

7.1 experiment of machine learning algorithm

An accuracy is preferred to compute the algorithm's performance analytically. The accuracy rate for a model is usually examine after the parameters for the model and calculates the percentage for the accuracy. Model prediction compares the true data for the measure of accuracy. A confusion matrix is a table used in machine learning to evaluate the performance of a classification algorithm. It provides a summary of the classification results by showing the count of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. These four elements are the fundamental metrics used to assess the performance of a classification model.



A screenshot of a computer

Description automatically generated

A screenshot of a computer error message

Description automatically generated

**CHAPTER – 8**

**Conclusion**

* The implementation of a robust spam mail prediction system contributes to an improved user experience. Users benefit from cleaner inboxes, reduced distractions, and increased confidence in the reliability of their email communications.
* Leveraging machine learning techniques, such as Support Vector Machines, Naive Bayes, and Neural Networks, demonstrates the potential for advanced pattern recognition and prediction capabilities. As these algorithms evolve, spam mail prediction systems can benefit from enhanced accuracy and adaptability.

**CHAPTER – 9**

**Reference**

1. Spam message filtering using python and machine learnig
   1. Online source :- <https://www.youtube.com/watch?v=YncZ0WwxyzU&t=5302s&ab_channel=CampusX>
2. kaggle for online dataset

:- <https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset>

1. Feature extraction of text data using Tfidf Vectorizer

:- <https://www.youtube.com/watch?v=Yo327cc818g&t=0s>

1. Overview of Pickle File:

<https://www.youtube.com/watch?v=6Q56r_fVqgw&ab_channel=Indently>

1. How to run code on Streamlit:

https://www.youtube.com/watch?v=Pe4gUSIxte8&ab\_channel=DatAInvestor

Thank you so much!!!