|  |  |  |
| --- | --- | --- |
| Name | Roll No | Class |
| Siddhant Sanjay Sathe | 50 | D15A |
| Arnav Santosh Sawant | 52 | D15A |
| Pranav Pramod Titambe | 60 | D15A |

**Topic:- Email Spam Classifier: A Machine Learning Approach**

**Introduction:-**   
Email spam, or junk mail, is a major issue, often containing phishing, malware, or fraud, posing risks to individuals and organizations. Machine learning offers an effective solution for automatically detecting and filtering spam. This report explores ML-based spam classifiers, covering relevant literature, methodology, system diagram, advantages, disadvantages, applications, and objectives.

**Research papers:-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Paper Name** | **Date** | **Abstract** | **Key Points** |
| **"A comparative study of machine learning methods for email spam detection"** | Dec, 2020 | This paper compares several machine learning algorithms, including Naive Bayes, Support Vector Machines (SVM), and Random Forests, for email spam detection. The performance of each algorithm is evaluated based on metrics such as accuracy, precision, and recall. | - Comparative analysis of multiple ML algorithms. - Performance evaluation using standard metrics. - Highlights the strengths and weaknesses of each algorithm for spam detection. |
| **"Spam Email Detection based on Machine Learning Techniques"** | May, 2023 | This study proposes a spam email detection system using machine learning techniques such as Naive Bayes and Support Vector Machine (SVM). The system is trained on a dataset of labeled emails and uses feature extraction techniques to identify relevant characteristics of spam emails. The results demonstrate the effectiveness of the proposed system. | - Feature extraction techniques for identifying spam characteristics. - Training and evaluation of ML models on labeled email datasets. - Focus on Naive Bayes and SVM algorithms. |
| **"An efficient email spam detection system using hybrid feature selection and machine learning techniques"** | June, 2024 | This paper introduces an efficient email spam detection system that combines hybrid feature selection and machine learning techniques. The proposed system uses a combination of filter and wrapper-based feature selection methods to select the most relevant features for spam detection. | - Hybrid feature selection methods for improved accuracy. - Focus on efficiency and reducing computational complexity. - Integration of feature selection with machine learning algorithms. |

**Methodology:-**   
· **Data Collection:** Gather a labeled email dataset (spam and ham), e.g., SpamAssassin or Enron-Spam.

· **Preprocessing:** Clean and normalize the text by:

* Removing HTML tags, special characters, and stop words.
* Converting to lowercase.
* Applying stemming or lemmatization.

· **Feature Extraction:** Convert emails into numerical features using:

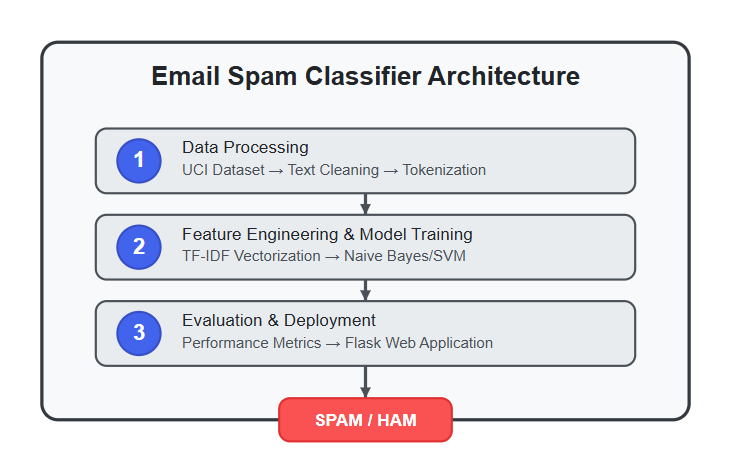
* **BoW:** Word frequency representation.
* **TF-IDF:** Weighing words by frequency and rarity.
* **N-grams:** Sequences of n words for context.
* **Header Features:** Detecting suspicious patterns.

· **Model Training:** Train ML models (e.g., Naive Bayes, SVM) using extracted features.

· **Evaluation:** Measure performance with:

* **Accuracy:** Overall correct classifications.
* **Precision:** Correct spam predictions out of all spam-labeled emails.
* **Recall:** Correct spam predictions out of all actual spam emails.
* **F1-score:** Harmonic mean of precision and recall.

· **Deployment:** Use the trained model to filter real-time emails.

**Diagram:-   
**

**Advantages:-**

* **High Accuracy**: Machine learning models can achieve high accuracy in detecting spam emails by learning complex patterns and relationships in the data.
* **Automation**: Automated spam filtering saves users time and effort by automatically classifying and filtering unwanted emails.

**Disadvantages:-**

* **Data Dependency:** The performance of machine learning models heavily depends on the quality and quantity of training data.
* **Computational Cost:** Training and deploying complex machine learning models can be computationally expensive.

**Applications**

* **Email Service Providers:** Use spam filters to protect users from unwanted and malicious emails.
* **Corporate Email Systems:** Implement spam filters to safeguard sensitive information and prevent phishing attacks.

**Objectives**

* To collect and preprocess a diverse dataset of email messages labeled as spam or ham.
* To extract relevant features from email content and headers using techniques like TF-IDF, and N-grams.
* To train and evaluate machine learning models, such as Naive Bayes, SVM, and Random Forest, for email spam classification.
* To optimize model parameters and feature selection methods to achieve high accuracy, precision, and recall.
* To develop a system that can effectively filter spam emails while minimizing false positives.