Email Spam Detection Using Logistic Regression

Abstract-

This project aims to classify email messages as either spam or ham (non-spam) using a machine learning approach. Logistic Regression is employed as the classification model, combined with feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF). The project demonstrates a high degree of accuracy, leveraging text preprocessing, feature engineering, and model evaluation to address the problem of spam detection. The final model achieves an accuracy of 98.3% on the test dataset, demonstrating its effectiveness in real-world applications.

1. Introduction-

With the increasing volume of unsolicited spam emails, it has become critical to develop efficient and scalable methods for filtering such emails. Traditional rule-based filtering approaches often struggle with the dynamic nature of spam. This project addresses the problem using a machine learning-based approach, specifically Logistic Regression, to automatically classify emails based on their content.

2. Dataset-

The dataset consists of 5,572 email messages, labeled as either "spam" or "ham." The "Message" column contains the body of the email, while the "Category" column contains the label, with 1 for spam and 0 for ham. The dataset is split into 80% training data and 20% test data.

3. Methodology-

3.1. Data Preprocessing-

Before training the model, the dataset is cleaned and prepared:
Handling Missing Values: Null values in the dataset were replaced with empty strings.
Label Encoding: Labels were encoded as 0 for "ham" and 1 for "spam."
Splitting Data: The dataset was divided into a training set (80%) and a test set (20%) using a random split.

3.2. Feature Extraction (TF-IDF Vectorization)-

Text data was converted into numerical features using the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF is effective in transforming textual data into a format suitable for machine learning models by quantifying the importance of words in each document relative to the entire dataset. This step ensures that common words are weighted less, while rare but important words are weighted higher.

3.3. Logistic Regression Model-

A Logistic Regression model with cross-validation (LogisticRegressionCV) was used to classify emails as spam or ham. Logistic Regression was chosen due to its simplicity, effectiveness, and

interpretability in binary classification problems. The model was trained on the TF-IDF transformed feature set.

3.4. Model Evaluation-

The model was evaluated using accuracy as the primary metric, comparing predictions with actual labels on both training and test datasets. The model achieved an accuracy of 100% on the training data and 98.3% on the test data, demonstrating its generalization capability.

4. Results and Discussion-

The Logistic Regression model performed exceptionally well in classifying emails as spam or ham. With an accuracy of 98.3% on the test data, the model has shown that it can effectively handle text classification tasks. The slight drop in accuracy from the training data to the test data suggests that the model generalizes well without overfitting. The TF-IDF method proved crucial for converting raw text into meaningful features.

5. Conclusion:

The project successfully implemented a machine learning-based solution to the problem of email spam detection. Using Logistic Regression and TF-IDF vectorization, the model achieved high accuracy on both training and test datasets. Future work could explore more sophisticated models like Naive Bayes or deep learning techniques, as well as additional feature engineering to further improve performance.

6. Future Work:

Possible future enhancements include:
- Experimenting with other classification algorithms like Naive Bayes, Support Vector Machines (SVM), or ensemble methods (Random Forest, XGBoost).
- Incorporating additional features such as email metadata (sender, subject line) for more accurate detection.
- Testing the model's performance on larger and more diverse datasets.
- Building a real-time spam detection system integrated with email clients.

7. References:

- **TF-IDF:** Term Frequency-Inverse Document Frequency
- Logistic Regression for Binary Classification
- Scikit-learn Documentation for Feature Extraction and Classification Models