Automated Email Spam Detection Using RPA and Machine Learning

Student: German Gnetov
ID: 11941513

Introduction

This project aims to automate the classification of text documents, representing emails, as spam or non-spam. I employ a machine learning model coupled with robotic process automation (RPA) to streamline this process. The project is designed such that it can be adapted for actual email processing in the future.

Project Environment and Setup

Tools and Software: Python 3.9, Scikit-learn for machine learning, RPA for automation.
Folder Structure:
emails/: Contains text documents simulating emails.
templates/: html code of my mini website.

Data Preparation

Data Source: The training dataset contains text column which represents real-life emails and label column with actual classification (spam or not).
Preprocessing: Text data is preprocessed using techniques like lowercasing, tokenization, and removal of stopwords to normalize the data.

Machine Learning Model

Model Used: An MLP Classifier was chosen for its efficacy in text classification tasks.
Training Process: The model was trained on a labeled dataset, split into 70% training and 30% testing.

RPA Implementation

Read Text Files: The RPA bot is programmed to sequentially access text documents from the folder.
Classification: Each text document is fed into the machine learning model to classify it as spam or non-spam.
Logging Results: The classification results are logged for each document for review.
Challenges: Integrating the Python-based model with the RPA required custom scripting to handle data exchange between the two systems. Especially problematic was the format of data input since the model expects strings as input.

Results and Model Evaluation

Performance Metrics: The model achieved an accuracy of 98%, with precision and recall metrics also indicating high reliability.
RPA Effectiveness: The RPA bot successfully automated the classification process, demonstrating potential for scaling up to handle actual email workflows.

Conclusion and Future Enhancements

Summary: The project successfully demonstrates the integration of machine learning and RPA to automate email spam detection.
Future Work: Future enhancements could include adapting the system to directly process emails from a mail server, and expanding the model to handle different languages and more complex email formats.