Email Classification for Support Team

1. Introduction

In today's digital landscape, emails are a primary mode of communication for both personal and professional use. However, with the rising volume of emails, it has become increasingly difficult to manually process and route messages efficiently. Moreover, emails often contain **Personally Identifiable Information (PII)** such as names, phone numbers, and email addresses, which must be handled with care to comply with data protection regulations like GDPR and HIPAA.

This project addresses two key challenges:

1. **Email Classification**: Automatically categorizing incoming emails into labels such as HR, Finance, Support, etc.
2. **PII Masking**: Detecting and anonymizing sensitive data in the email body to preserve user privacy.

The end goal is to build a deployable system (via API) that provides both classification and PII masking in real-time.

2. Approach

2.1 Email Classification

To classify email content into predefined categories, the following NLP pipeline was implemented:

- **Preprocessing**: Lowercasing, stop-word removal, punctuation stripping, and tokenization.
- **Text Vectorization**: TF-IDF (Term Frequency-Inverse Document Frequency) was used to convert text into numerical vectors representing word importance.
- **Model**: Support Vector Machines.

2.2 PII Masking

PII detection and anonymization involved two techniques:

- **Regex-based Masking**: Used for patterns like email addresses and phone numbers.

3. Model Selection and Training

After preparing a labeled dataset of emails (sourced and/or simulated), the classification model was trained using scikit-learn.

Training Details:

- **Data split**: 80% train, 20% test

- **Vectorizer**: TF-IDF with unigrams and bigrams
- **Model**:Support Vector Machines
- **Metrics used**: Accuracy, Precision, Recall, F1-score

The model generalized well and showed consistent performance across categories.

4. Challenges and Solutions

🔷 Challenge 1: Inconsistent Email Formats

Emails varied greatly in format — some were structured, others informal and some of them in different language .

- **Solution**: Created robust regex expressions and email with differnet language catagorised as other.

🔷 Challenge 2: Ambiguity in Short Emails

Short emails with vague content were difficult to classify accurately.

- **Solution**: Augmented training data with more representative samples and introduced bigrams in the TF-IDF vectorizer for better context.

🔷 Challenge 4: Deployment Issues

Initial deployment on Hugging Face Spaces faced timeouts and missing dependencies.

- **Solution**: Added a `requirements.txt`, optimized the code, and ensured proper binding to ports/environment variables.

Conclusion

This project successfully developed a system that can automatically classify emails and mask sensitive personal information. By combining machine learning for classification and NLP techniques for PII detection, the solution provides an efficient and privacy-conscious way to process emails. The system is also deployed as an API, making it easy to integrate into real-world applications for secure and automated email handling.