Digital Empowerment Network

Data Science

Week 03

Email Spam Classification

Mentor : Ali Mohiuddin Khan

Task 3: Email Spam Classification — Mastering Text Data in Machine Learning

Introduction

In this task, your objective is to build a classification model that can distinguish between spam and ham (not spam) emails using a labeled dataset of messages.

This is your first Natural Language Processing (NLP) task — working with text data — and it introduces you to classification, an essential ML concept.

Step-by-Step Guide to Approach Task 3

1. Data Cleaning and Preprocessing

Goal: Prepare the raw text data for model training.

Important Steps:
- Remove punctuation, stopwords (like "the", "is", "and"), and unnecessary characters.
- Convert text to lowercase.
- Tokenize the text (split into words).
- Optionally, apply stemming or lemmatization.

Tools to use: nltk, re, sklearn, pandas

Why it matters: Raw text must be cleaned and structured before the model can understand it.

2. Vectorization (Converting Text to Numbers)

Goal: Convert text into a numerical format that models can work with.

Techniques:
- Bag of Words using CountVectorizer
- TF-IDF Vectorizer

Tools to use: Scikit-learn's CountVectorizer and TfidfVectorizer

Why it matters: Machine learning models cannot process text directly — words must be converted into numerical vectors.

3. Model Building — Binary Classification

Goal: Build a model that classifies each email as spam or ham (not spam).

Algorithms to try:

- Multinomial Naive Bayes (recommended for text)
- Logistic Regression
- Support Vector Machine (SVM)

Tools to use: Scikit-learn classifiers

Why it matters: You are building a model for a real-world problem. Spam detection is a common and important classification task in industry.

4. Model Evaluation

Goal: Understand how well your model is performing.

Evaluation Metrics:
- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

Tools to use: classification_report and confusion_matrix from Scikit-learn
Why it matters: In spam detection, precision and recall are especially important — the goal is to reduce false positives and false negatives.

Deadline 6 – August -2025