WEEK 3 SUBMISSION

Email Spam Classification

Muhamm KAIF UR REHMAN
STU-DS-251-232

Email Spam Classification Report

Digital Empowerment Network – Week 03 Task

Mentor: Ali Mohiuddin Khan

Introduction

This project aimed to build a model to classify emails as spam or ham using Natural Language Processing (NLP) techniques. The models were trained using the UCI SMS Spam Collection Dataset.

Step 1: Preprocessing

- Lowercased all text
- Removed punctuation and numbers
- Removed stopwords
- Applied Porter Stemming

Step 2: Vectorization

Used **TF-IDF Vectorizer** to transform text into numerical format.

Step 3: Model Evaluation

1. Multinomial Naive Bayes

Accuracy: 0.9659

Classification Report:

Label	Precision	Recall	F1-Score	Support
Ham (0)	0.96	1.00	0.98	965
Spam (1)	1.00	0.75	0.85	150
Accuracy			0.97	1115
Macro Avg	0.98	0.87	0.92	

Weighted	0.97	0.97	0.96	

Confusion Matrix:

	Predicted Ham	**Predicted Spam**
Actual Ham	965	0
Actual Spam	38	112

2. Logistic Regression

Accuracy: 0.9552

Classification Report:

Label	**Precision**	**Recall**	**F1-Score**	**Support**
Ham (0)	0.95	1.00	0.97	965
Spam (1)	0.96	0.69	0.81	150
Accuracy			0.96	1115

Confusion Matrix:

	Predicted Ham	**Predicted Spam**
Actual Ham	964	1
Actual Spam	24	126

✅ **Summary Comparison**

Model	**Accuracy**	**Precision (Spam)**	**Recall (Spam)**	**F1-Score (Spam)**
Multinomial Naive Bayes	96.59%	1.00	0.75	0.85
Logistic Regression	95.52%	0.96	0.69	0.81