Spam Email Detection using Machine Learning & NLP

Overview

This project is an AI - powered Spam Email Detection system that classifies emails as Spam or Not Spam using Natural Language Processing (NLP) and Machine Learning algorithms. The system performs text preprocessing, TF - IDF feature extraction, model training, evaluation, and real - time prediction through an interactive interface.

Features

- NLP - based spam classification
- TF - IDF vectorization for feature engineering
- Multiple ML models: Na ï ve Bayes, Logistic Regression, Random Forest
- Model comparison and evaluation metrics
- Real - time email prediction interface

Tech Stack

Programming: Python
Libraries: Scikit - Learn, Pandas, NumPy, NLTK
ML Techniques: TF - IDF, Classification Algorithms, Model Evaluation
Optional UI: Streamlit
Version Control: Git & GitHub

Machine Learning Workflow

1. Data Collection – Spam/Ham email dataset
2. Preprocessing – Cleaning, tokenization, stopword removal
3. Feature Engineering – TF - IDF vectorization
4. Model Training – NB, LR, RF classifiers
5. Evaluation – Accuracy, Precision, Recall, F1 - score
6. Deployment – Real - time prediction via Streamlit/FastAPI

Results

Best accuracy achieved using Logistic Regression (~98%).

How to Run

1. Clone the repository
2. Install requirements using pip
3. Train the model
4. Launch Streamlit app for real - time prediction

Author

Nagendra Babu Kondameda
MS in Information Technology Management
Java Developer | SDET | AI/ML Enthusiast
Email: Nagendrababu9152@gmail.com

Model Accuracy Comparison

Model	Accuracy
Na ï ve Bayes	97%
Logistic Regression	98%
Random Forest	96%