Spam Email Detection Using NLP and Machine Learning

Abstract:
Spam emails are a significant problem for email users and businesses, leading to unnecessary clutter and potential security risks. This project focuses on developing an **email spam detection system** using **Natural Language Processing (NLP)** and **machine learning** techniques to classify emails as either spam or legitimate.

Methodology:
The system will preprocess email content using **text cleaning**, **tokenization**, and **stopword removal**. **Text vectorization techniques** such as **TF-IDF** or **word embeddings** will be employed to convert the email text into numerical features. A combination of **machine learning models** like **Naive Bayes**, **Logistic Regression**, and **Random Forest** will be trained on labeled datasets containing both spam and non-spam emails. The system will also use **feature extraction** techniques, such as email metadata analysis (e.g., sender, subject line) to enhance classification accuracy.

Outcome:
The outcome will be a robust spam email detection system capable of accurately classifying emails, preventing spam from cluttering inboxes. The system will help reduce the risk of phishing attacks and malicious content, enhancing email security for users and organizations.