**ABSTRACT**  
 **NLP-Based Phishing Email Detection**

This project presents an effective email classification system that utilizes Natural Language Processing (NLP) and machine learning to detect phishing or spam emails. The workflow begins with data preprocessing, where missing values in the subject and body fields are handled, and labels are standardized. Text data is cleaned by removing punctuation, stopwords, and case inconsistencies to retain only meaningful information. Feature extraction includes both TF-IDF-based textual features and engineered features such as the number of hyperlinks, uppercase words, special characters, email length, and word count. These features are combined into a unified representation using sparse matrix stacking. A Logistic Regression model is trained on this feature set and evaluated using accuracy, precision, recall, and F1-score metrics. A prediction pipeline is implemented to classify new emails based on their subject and body content. The model effectively distinguishes between legitimate (ham) and spam emails, demonstrating strong performance in real-world scenarios like email clients or corporate email filters. Future improvements may involve advanced algorithms or enhanced feature engineering for greater accuracy and adaptability.