**Spam Detection**

**Group Members**

Md Tajriyaan Chowdhury, Rafid Hasan, Ali Rizvi

**1. Introduction**

Spam emails disrupt communication by sending unsolicited and often malicious messages, threatening user security and productivity. Detecting spam is a critical NLP application that aims to safeguard individuals and organizations from fraud, malware, and information overload. The objective of this project is to evaluate and compare the performance of multiple models to determine which achieves the highest metrics and reliability in identifying spam. We highlight the evolving nature of spam, including image-based, obfuscated, and context-aware attacks, emphasising the real-world impact of effective detection.

**2. Dataset Description**

We utilize the [Enron Spam Dataset](https://www2.aueb.gr/users/ion/data/enron-spam/), which contains 33,716 emails consisting of 17,171 spam and 16,545 ham messages. This includes subjects, message bodies, labels, and dates. The dataset is publicly available and well-balanced, enabling rigorous supervised learning and facilitating benchmarking against prior studies. All data usage complies with licensing and ethical standards.

**3. Problem Statement and Task**

Given the text content of an email, either its subject and/or body. The task is to classify it as "spam" or "ham" (non-spam). The primary challenge lies in detecting diverse spam tactics while minimizing false positives on legitimate emails. Following classification, a comparative analysis across different models will determine which approach is most suitable for this detection task.

**4. Related Work**

Past research on spam filtering has heavily utilized Naïve Bayes and Support Vector Machines (SVMs). More recently, transformer-based models such as BERT have shown strong performance in semantic classification. Our work extends these efforts by directly comparing classical models with modern machine learning techniques, aiming to replicate and exceed published benchmarks.

**5. Proposed Approach**

We plan to use six different machine learning models: Naïve Bayes, Logistic Regression, BERT, SVM, Random Forest, and K-Nearest Neighbors.

**6. Evaluation Metrics**

Our primary evaluation metric will be the macro-average F1 score, which fairly balances precision and recall for both spam and ham classes. Additionally, we will report accuracy, precision, and recall as secondary metrics. The F1 score is ideal given the occasional class imbalance and the need to avoid both false positives and negatives.

**7. Ethical and Responsible NLP Considerations**

We will strictly comply with dataset licensing and privacy standards, ensuring the ethical handling of personal identifiers in emails. Our analysis will include assessment of bias in the dataset composition, model fairness, and the potential for societal harm. We will also discuss misclassification risks, such as blocking legitimate emails or missing spam threats, ensuring alignment with the ACL ARR Responsible NLP Research Checklist.