**Spam Detection Project Proposal**

**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, aiming to safeguard individuals and organizations from fraud, malware, and information overload. The objective of the 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, emphasizing 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 (17,171 spam and 16,545 ham), including subjects, message, label and date. This is a publicly available, well-balanced dataset which enables rigorous supervised learning and facilitates benchmarking against prior studies. All data usage complies with licensing and ethical standards.

4. Problem Statement and Task

Given the text content of an email (subject and/or body), classify it as "spam" or "ham” (non-spam). The primary challenge lies in detecting diverse spam tactics while minimizing false positives on legitimate emails. Afterwards, across different models we must analyze which one is the most suitable for this detection technique and do a comparative analysis.

**5. Related Work**

Past research on spam filtering has heavily used Naïve Bayes and SVMs. More recently, transformer-based models like BERT have shown strong performance in semantic classification. Our work extends this by directly comparing classical models with modern machine learning techniques, aiming to replicate and exceed published benchmarks.

**6. Proposed Approach**

We are choosing six different machine learning models - Naïve Bayes, Logistic Regression, BERT, SVM, Random Forest, and K-Nearest Neighbors.

**8. Evaluation Metrics**

Primary: Macro-average F1 score (to fairly balance precision and recall for both spam and ham)

Secondary: Accuracy, precision, recall

F1 is ideal given occasional class imbalance and the need to avoid both false positives and negatives.

**9. Ethical and Responsible NLP Considerations**

Strict compliance with dataset licensing and privacy standards

Ethical handling of personal identifiers in emails

Analysis of bias in dataset composition, model fairness, and potential for societal harm

Discussion of misclassification risks (e.g., blocking legitimate emails, missing spam threats)

Project aligns with the ACL ARR Responsible NLP Research Checklist