### ****Spam Email Classification Project Updates****

Our group has completed data preprocessing, feature engineering, and initial modeling. We cleaned the dataset, checked for missing values and duplicates, and engineered features such as email length, subject length, number of links, presence of urgency/financial keywords, capitalization, and sentiment. We also applied text cleaning and lemmatization to the email body, then extracted TF-IDF features using the top 1000 1–2 gram terms.

For exploration, we compared spam and ham emails. Spam messages tend to have shorter subjects, longer bodies, more links, and use more urgency words. Domain analysis also showed clear differences in sender patterns between spam and ham emails.

We trained Naive Bayes, Logistic Regression, and Random Forest models using our engineered features. Logistic Regression and Random Forest achieved perfect performance (100% accuracy, precision, recall, and AUC). Feature importance plots showed body\_length and email\_length as the strongest predictors, with urgency and sentiment also contributing.

Overall, our features effectively separate spam from ham, and our models perform very well. If it’s convenient, we’d really appreciate it if you could help us run the notebook code on your end — we’re a bit unsure whether all sections are working as expected, especially during modeling and evaluation. Thank you!