# EMAIL SPAM CLASSIFIER

# Motivation:

The project aims to develop an email spam classifier that filters out all unwanted and harmful emails from user's inboxes. Spam emails generally contain scams and malicious content that can compromise personal and organizational security.

In this project, we build an accurate automated spam detection system which aims to enhance email security without any manual effort of identifying spam and improve user experience by ensuring that relevant emails are prioritized and spam messages are minimized

# Problem Statement:

Email has become a vital medium for communication in both personal and professional aspects of life but this leads to rising spam containing malware and scams that threaten users. Traditional rule-based filtering systems cannot help with the evolving tactics used by spammers.

The problem is to make a well-developed data-driven email spam classifier using NLP techniques which helps in automatically detecting and filtering emails with high accuracy. This classifier should handle diverse and evolving spam tactics and ensuring to minimize errors, it should also ensure that important emails are not mistakenly filtered as spam messages. This solution should be efficient for real-time deployment and flexible to handle evolving spam patterns.

# Proposed Pipeline:

- DATA COLLECTION:

>We use spam assassin as our labelled email datasets, this includes both spam and valid emails.

- DATA PREPROCESSING: This includes

>**Cleaning**: We remove all unwanted characters like email metadata, HTML tags that don't contribute in classification.

>**Normalisation**: Here we apply techniques like Tokenisation, Lemmatization and Removal of stop words. Which converts Data into Standardized form.

>**Feature Engineering:** We use relevant features like text content features (word count,length),sender information and attachment types.

>**Feature Extraction**: To convert the preprocessed text into numerical representation as most ML algorthims require numerical input. Here,we use TF-IDF method .it reduces all common words that are of no use.

- MODEL SELECTION AND TRAINING**:**

We selected logistic regression as our NLP algorithm model for this project. cross validation is done. class balancing is done incase of  inbalanced dataset ,we use SMOTE as our method here.

- MODEL EVALUTION:

>Assessing Model Performance Through:

Accuracy, Precision, Recall, f1-score.

here, in some cases valid emails are classifieed into spam and spam emails are undetected.to minimise these type of errors model evaluation is done.

- DEPLOYMENT AND TESTING:

  >Implement the trained model for real-time spam detection within an email system initially by using local deployment and then cloud platform.Validate the model on a separate test set to assess performance and robustness

- DOCUMENTATION:

>we record all steps ,methodoligies and challanges performed.

# Timeline:

- Week 1: Project Planning and Research
- Week 2: Data Collection and Preparation
- Week 3: Data Preprocessing and Feature Engineering
- Week 4: Model Selection and Training
- Week 5: Model Evaluation and Tuning
- Week 6: Deployment Preparation
- Week 7: Deployment
- Week 8: Testing, Documentation, and Final Presentation

# Expected Outcome:

- By the end of the project,We create a working model which accurately classifies spam and legimate emails.a RESTful API is created that allows users to submit a email and recieve spam classification in real time.