EMAIL SPAM DETECTION

Name: Jayachandiran V
NMid:au513521114007
Department: Mechanical Engineerin
College:AMCET

CONTENT

- OBJECTIVE
- INTRODUCTION
- SPAM AS A PROBLEM
- OBTAINING EMAIL THROUGH VARIOUS METHODS
- LIFECYCLE OF SPAM
- TYPES OF SPAM FILTERS
- FLOWCART OF PROCESSING
- DOCUMENT PREPROCESSING
- SCOPE OF THE PROJECT

OBJECTIVE

- To give knowledge to the user about fake emails and relevant emails.
- To classify the mail as spam or ham.

INTRODUCTION

What is SPAM?

- **Spam also called as Unsolicited Commercial Email(UCE)**
- **Involves sending message by email to numerous recipients at the same time (Mass Emailing)**
- **Grew exponentially since 1990**
- **80% of all spam is sent by less than 200 spammers.**

Purpose of SPAM

- **Advertisement**
- **Pyramid Schemes(Multi Level Marketing)**
- **Giveaways**
- **Chain Letters**
- **Political Email**
- **Stock Market Advice**

SPAM AS A PROBLEM

- Consumes computing resources and time.
- Reduces the effectiveness of legitimate advertising
- Cost Shifting

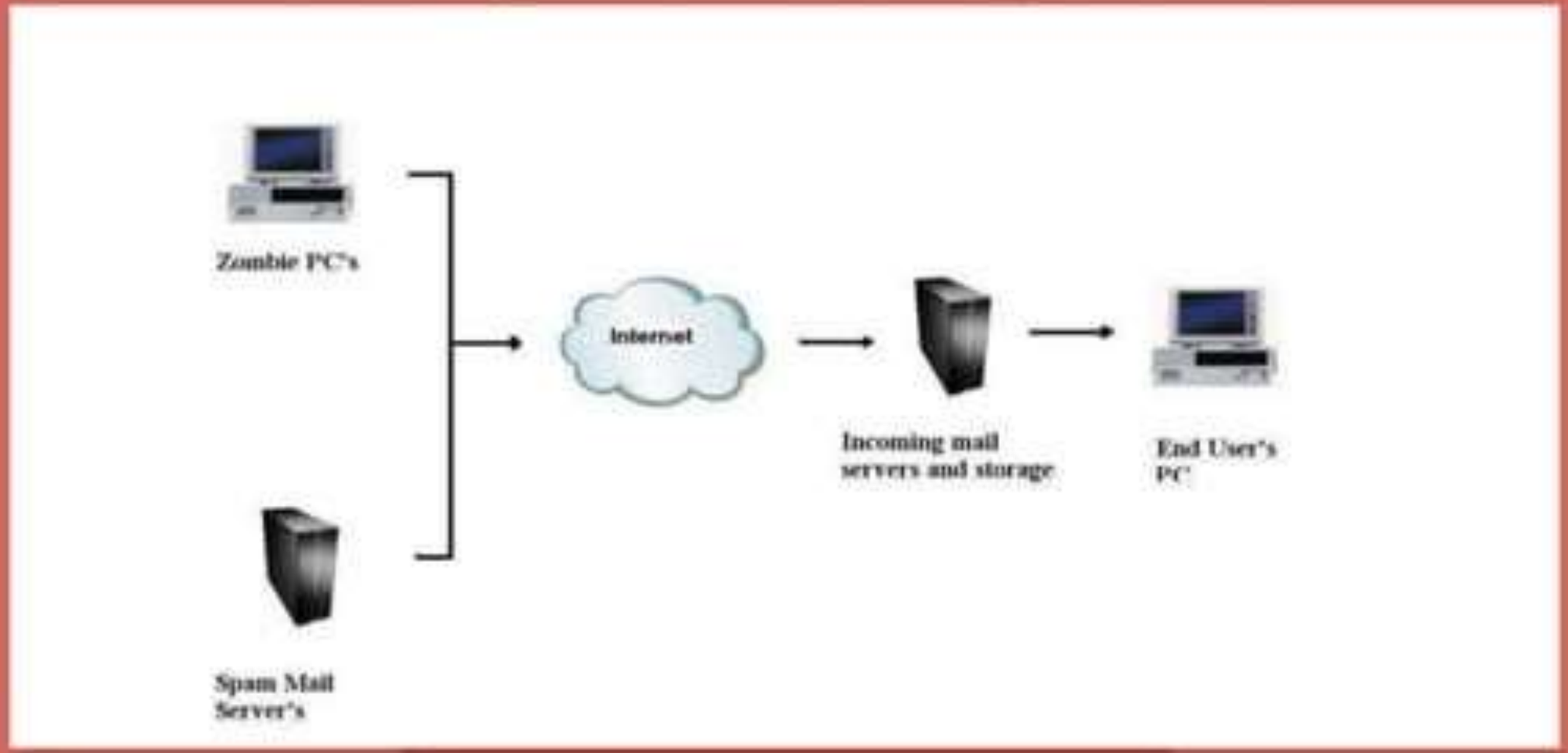
- Fraud
- Identity Theft
- Consumer Perception
- Global Implications

OBTAINING EMAIL THROUGH VARIOUS METHODS

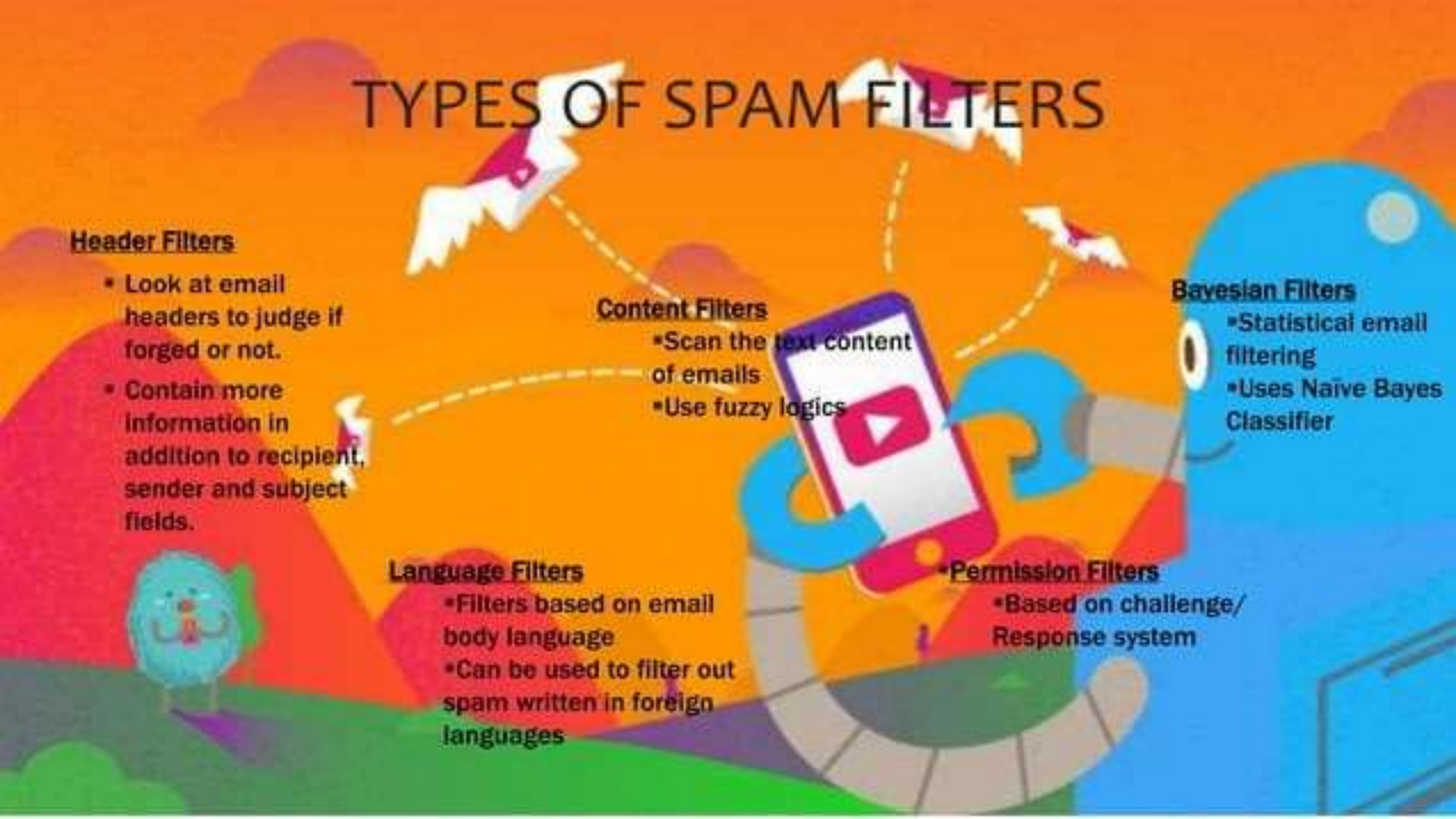
- Purchasing/ Trading lists with other spammers
- Bots
- Directory harvest attack

- Free Product or Services requiring valid email address
- News bulletins/Forums

LIFECYCLE OF A SPAM



TYPES OF SPAM FILTERS



Header Filters

- Look at email headers to judge if forged or not.
- Contain more information in addition to recipient, sender and subject fields.

Content Filters

- Scan the text content of emails
- Use fuzzy logics

Bayesian Filters

- Statistical email filtering
- Uses Naïve Bayes Classifier

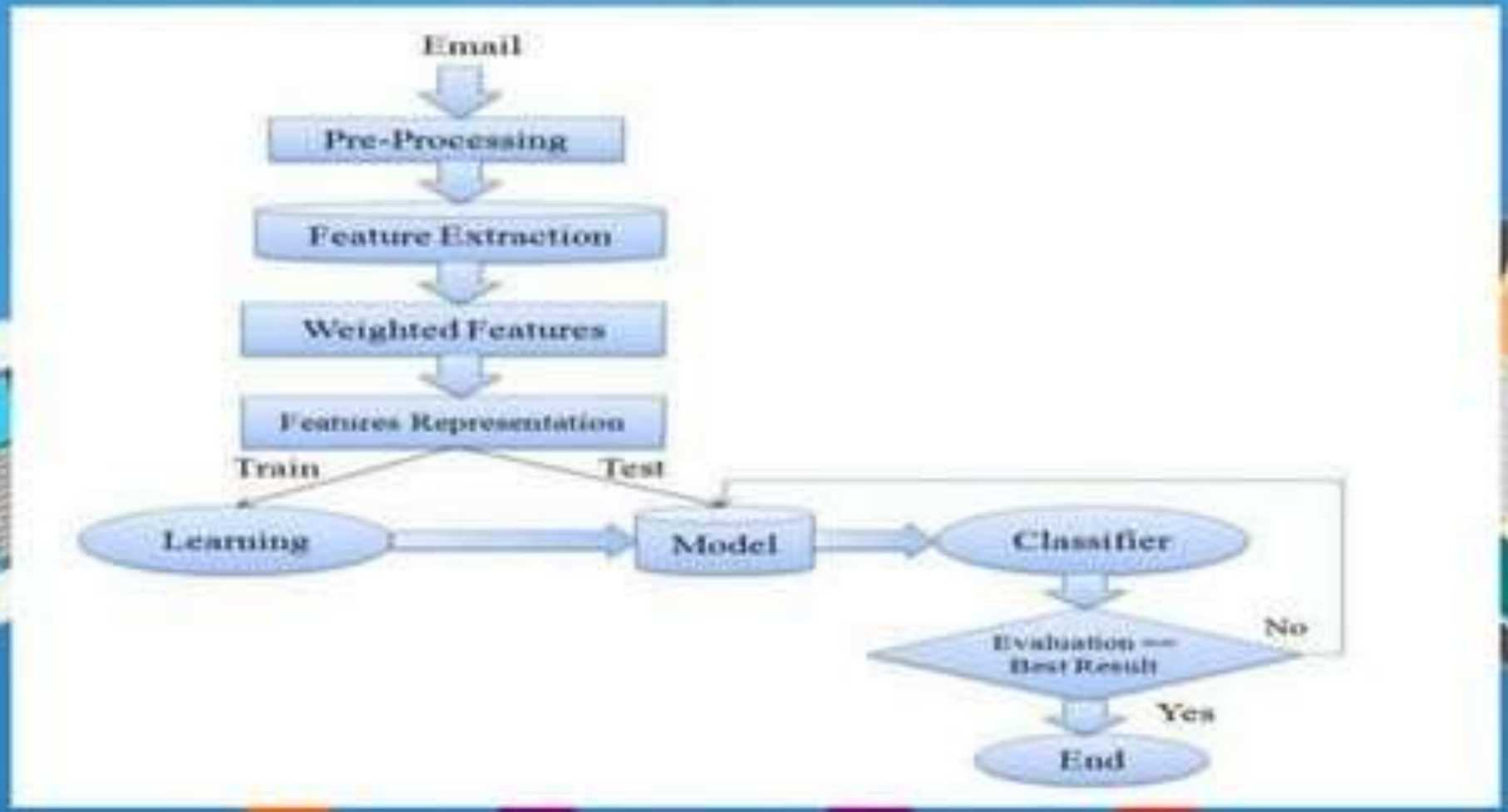
Language Filters

- Filters based on email body language
- Can be used to filter out spam written in foreign languages

Permission Filters

- Based on challenge/ Response system

FLOWCHART OF PROCESSING



DOCUMENT PREPROCESSING

Tokenization

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.

Lemmatization

Lemmatization linguistics, is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item.

Removal of Stop Words

Sometimes, the extremely common word which would appear to be of very little value in helping select documents matching user need are excluded from the vocabulary entirely.

SCOPE OF THE PROJECT

- It provides sensitivity to the client and adapts well to the future spam techniques.
- It considers a complete message instead of single words with respect to its organization.
- It increases Security and Control.
- It reduces IT Administration Costs.
- It also reduces Network Resource Costs.

THANK YOU