SPAM MAIL DETECTION

S Abhishek	AM.EN.U4CSE19147
Navneet Kumar Singh	AM.EN.U4CSE19138
Harsha Sathish	AM.EN.U4CSE19123
Arvind Kumar K	AM.EN.U4CSE19109

ABSTRACT

Spam email is one of the most demanding and troublesome internet issues in today's world of communication and technology. Sending malicious link through spam emails which can harm our system and can also seek in into your system. Spam emails not only influence the organizations financially but also exasperate the individual email user. So, it is needed to Identify those spam mails which are fraud and this project will identify those spam by using techniques of machine learning, where it applies algorithms on our data sets and the best algorithm is selected for the email spam detection having best precision and accuracy.

Ill Posed Problem:

- I need to classify the spam and non-spam mails.

Well Posed Problem:

- **Task** – Classifying emails as spam or not
- **Performance Measure** – The fraction of emails accurately classified as spam or not spam
- **Experience** – Observing you label emails as spam or not spam.

INTRODUCTION

Motivation

- In our day-to-day life, Spam Emails are considered to be annoying and repetitive, which is solely send for the purpose of advertisement and brand promotion.
- People are using them for illegal and unethical conducts, phishing and fraud.
- Sending malicious link through spam emails which can harm our system and can also seek in into your system.
- Creating a fake profile and email account is much easy for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds.
- Although we block such emails, it is of no use as spam emails are still prevalent.
- Thus, we need to build a robust real-time email spam classifier that can efficiently and correctly flag the incoming mail spam, as either a spam or ham (non – spam) email.

Benefits

- Spam filters can provide a great firewall to the spam emails which can be carriers of dangerous computer viruses.

- Spam filter that blocks spam emails from reaching the inbox can save all important data.
- Spam filters also saves time. Business employees do not have to go through numerous emails to decide which ones are spams, as sometimes that can be hard to decide. The time saved can be used to increase productivity.
- Spam filters can help keep a company maintains its reputation. They can block viruses from reaching consumers data and prevent any spam mail accidentally being forwarded to consumers.
- Spam filters protect the servers from being overloaded with non-essential emails, and the worse problem of being infected with spam software that may turn them into spam servers themselves.

Solution Use

- Private companies, who have their own email servers, want their data to be more secure. In such cases, email spam classification solutions can be provided to them.
- Also, Employees of the company need not go through each and every email, and can sort out the spam emails from the list. Thus, the time saved can be used to increase productivity

- While this may sound like a straightforward task, it can be a challenge for filters that are not constantly updated according to the most recent spam techniques and senders.
- Spammers may change the address from which emails come or the wording inside the header or body to bypass out-of-date spam filters.
- This can be effective if the spam filter is not updated with the correct information on a regular basis.
- It is important to make sure that the spam filter has adequate spam intelligence.
- If it does, it can block hundreds or thousands of spam emails every month.
- Spam filters are available in the form of software, hosted, or an on-premise appliance.

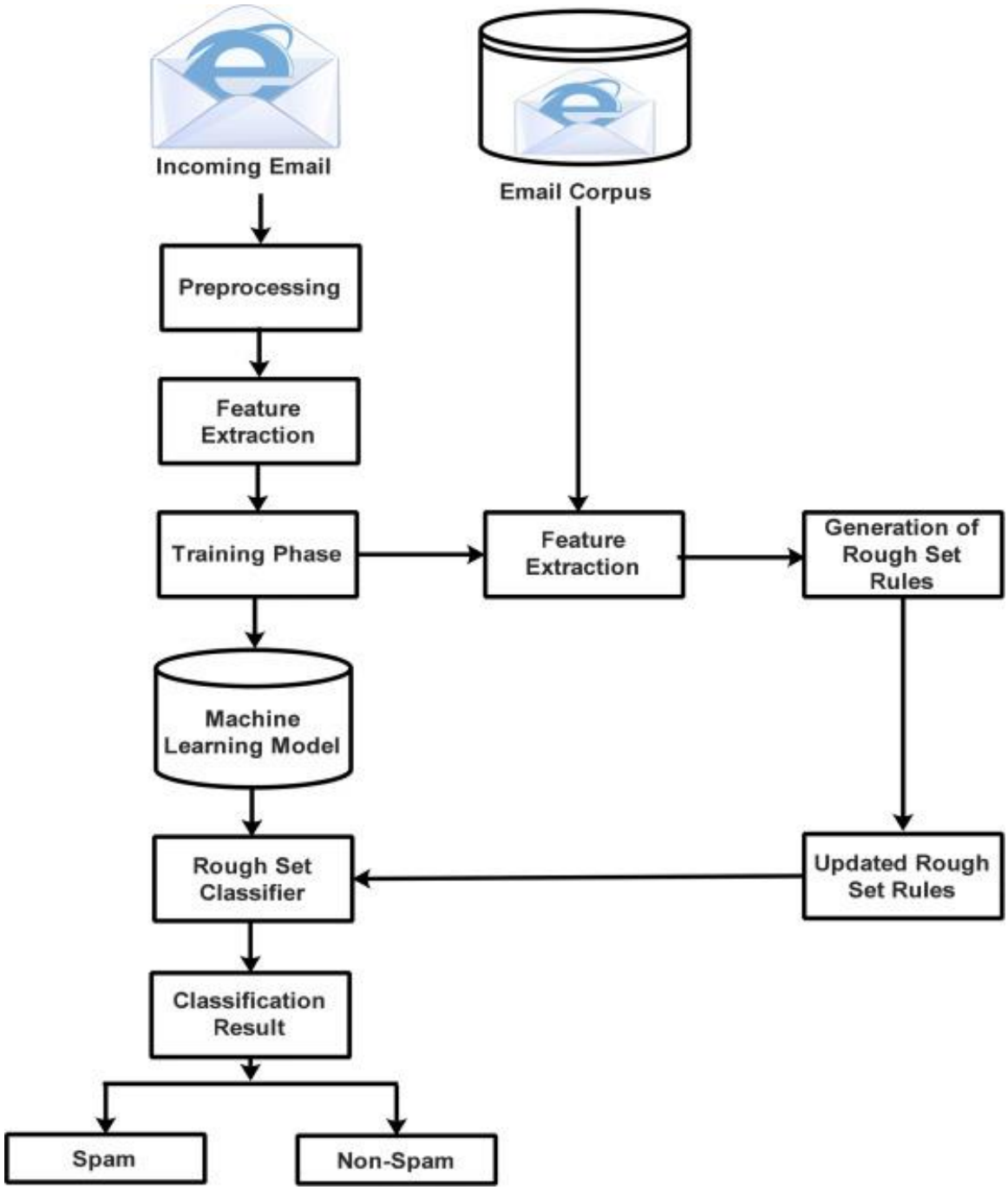
Functional Requirements

- The main function of this project is to clarify the e-mails which is done by first taking out the feature vector extraction which involves first taking out whether the word is a spam or not.

Non-Functional Requirements

- Ensures high availability of email datasets.
- User should get the results as fast as possible.

- It should be easy to use (ie) user is just required to type the words and click, then the result is displayed or the user is required to enter a pair of reasonable sentences.



DATASET FINALIZATION

Data Set 1	Data Set 2	Data Set 3	Data Set 4

- These datasets consist of emails sent mostly by the senior management of the Enron Corporation which contains most of the words or phrases that are particularly common in spam e-mails which are unprocessed/Unorganized.
- The dataset consists of 30207 emails of which 16545 emails are labelled as ham and 13662 emails are labelled as spam.
- Before using the data set for pre-processing it has to be organized with only useful information.
- In this experiment we are using a processed version of this dataset specifically made for spam and ham classification.

Features in the datasets

- There are around 6000 entries in each dataset approximately.
- There are **4** attributes in the data set.
    - Text
    - Spam
    - Length
    - Clean Text

- **Text**
    - Content of email from various sources.
    - This attribute contains the contents of email from various organizations along with the date, time, subject and the message.
- **Spam**
    - This attribute contains the classification of the emails whether it is spam/ham.
        - **1** – spam
        - **0** – ham
    - Using the key words related to spam mails, the received mails are classified as spam and not spam.
- **Length**
    - This attribute contains the count of characters in the email.
- **Clean Text**
    - This attribute contains the processed text, after removing all unnecessary characters like ": ," etc.
- The Enron Email Dataset was used in Corpus Linguistics and language analysis for email search and Expert search.

- Tech Giant use this data set to analyse the data from email analytics dashboard and also to compare it to the goals and the KPIs the company has set.
- Social media marketing companies use this data set to improve their email marketing results.

Assumptions

- Some of the datasets only has the content of the mail, and its not classified into spam and not spam (ie) Missing of attributes that are used to specify whether the mail belongs to spam or non spam.
- All the available datasets are not updated for a long period and thus they are not capable of identifying the most recent spam techniques and senders.
- We assume that, because of the non updation of the dataset over a period of time, we may miss all new spam keywords which is used to classify the spam and non-spam mails.