# Spam Mail Prediction

#### **Introduction**

In the era of information technology, information sharing has become very easy and fast. Many platforms are available for users to share information anywhere across the world. Among all information sharing mediums, email is the simplest, cheapest, and the most rapid method of information sharing worldwide. But, due to their simplicity, emails are vulnerable to different kinds of attacks, and the most common and dangerous one is spam. No one wants to receive emails not related to their interest because they waste receivers’ time and resources. Besides, these emails can have malicious content hidden in the form of attachments or URLs that may lead to the host system’s security breaches. Spam is any irrelevant and unwanted message or email sent by the attacker to a significant number of recipients by using emails or any other medium of information sharing. So, it requires an immense demand for the security of the email system. Spam emails may carry viruses, rats, and Trojans. Attackers mostly use this technique for luring users towards online services. They may send spam emails that contain attachments with the multiple-file extension, packed URLs that lead the user to malicious and spamming websites and end up with some sort of data or financial fraud and identify theft. Many email providers allow their users to make keywords base rules that automatically filter emails. Still, this approach is not very useful because it is difficult, and users do not want to customize their emails, due to which spammers attack their email accounts.

In the last few decades, Internet of things (IoT) has become a part of modern life and is growing rapidly. IoT has become an essential component of smart cities. There are a lot of IoT-based social media platforms and applications. Due to the emergence of IoT, spamming problems are increasing at a high rate. The researchers proposed various spam detection methods to detect and filter spam and spammers. Mainly, the existing spam detection methods are divided into two types: behaviour pattern-based approaches and semantic pattern-based approaches. These approaches have their limitations and drawbacks. There has been significant growth in spam emails, along with the rise of the Internet and communication around the globe . Spams are generated from any location of the world with the Internet’s help by hiding the attacker’s identity. There are a plenty of antispam tools and techniques, but the spam rate is still very high. The most dangerous spams are malicious emails containing links to malicious websites that can harm the victim’s data. Spam emails can also slow down the server response by filling up the memory or capacity of servers. To accurately detect spam emails and avoid the rising email spam issues, every organization carefully evaluates the available tools to tackle spam in their environment. Some famous mechanisms to identify and analyze the incoming emails for spam detection are Whitelist/Blacklist , mail header analysis, keyword checking, etc.

#### **Methodology**

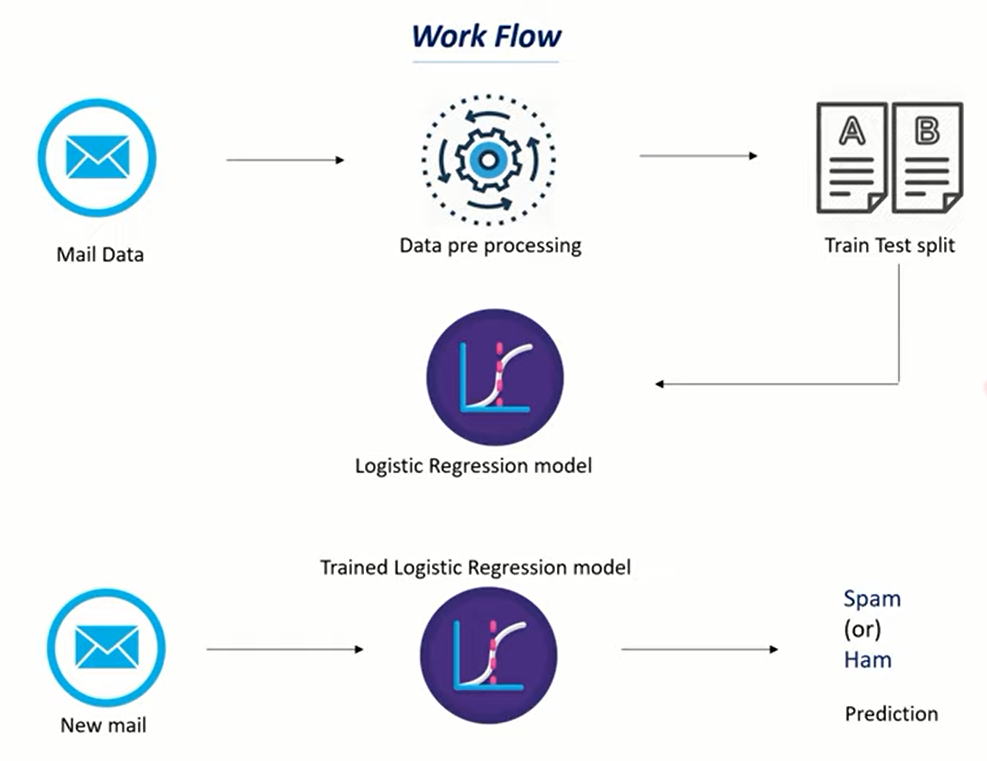
In the following sections, creating of dataset, training of learning models, and data preprocessing are explained.



##### **Data Preprocessing**

In machine learning (ML), the preprocessing phrase refers to organizing and managing of raw data before using it to train and test different learning models. In simplistic words, preprocessing is a ML data mining approach that turns raw data into a usable and resourceful structure.

The very first step in the construction of a ML model is preprocessing, in which data from the actual world, typically incomplete, imprecise, and inaccurate owing to flaws and deficient, is morphed into a precise, accurate, and usable input variables and trends



###### **Feature Extraction**

Feature extraction is the process of converting a large raw dataset into a more manageable format. Any variable, attribute, or class can be extracted from the dataset during this step, depending on the original dataset .

Feature extraction is a crucial step in training of the model, which helps in producing more reliable and accurate results. During the feature extraction process, out of the possible many attributes, the method of selecting some key variables that properly characterize data is called feature selection . The model is then constructed using these selected attributes or variables. If feature selection is performed properly, in return, the model construction will take less time.

**Logistic Regression Model**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the **Supervised Learning technique**. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**. Logistic Regression is much similar to the Linear Regression except that how they are used.

Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

#### **Conclusion**

In this study, we reviewed [machine learning approaches](https://www.sciencedirect.com/topics/computer-science/machine-learning-approach) and their application to the field of spam filtering. A review of the state of the art algorithms been applied for [classification](https://www.sciencedirect.com/topics/computer-science/classification) of messages as either spam or ham is provided. The evolution of spam messages over the years to evade filters was examined. The basic architecture of email spam filter and the processes involved in filtering spam emails were looked into. The study uses machine learning algorithms to detect them. We have to use Logistic Regression Model here. In the study, a translated emails dataset including spam and ham emails is generated from Kaggle. Accuracy, precision, F-measure, and model loss are used as comparative measures to examine performance. In addition, more recent artificial intelligent approaches may also be considered to detect spams.