**Machine learning for email spam filtering: review, approaches and open research problems**

**ABSTRACT**

In recent times, unwanted commercial bulk emails called spam has become a huge problem on the internet. Machine learning methods of recent are being used to successfully detect and filter spam emails. We present a systematic review of some of the popular machine learning based email spam filtering approaches. Our review covers survey of the important concepts, attempts, efficiency, and the research trend in spam filtering. The preliminary discussion in the study background examines the applications of machine learning techniques to the email spam filtering process of the leading internet service providers (ISPs) like Gmail, Yahoo and Outlook emails spam filters. Discussion on general email spam filtering process, and the various efforts by different researchers in combating spam through the use machine learning techniques was done. Our review compares the strengths and drawbacks of existing machine learning approaches and the open research problems in spam filtering.

**EXISTING SYSTEM**

The person sending the spam messages is referred to as the spammer. Such a person gathers email addresses from different websites, chatrooms, and viruses. Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time. In all existing systems doesnot find the spam mails effectively. It is also resulted to untold financial loss to many users.

**Disadvantages of Existing System:**

1. Less Security.

**PROPOSED SYSTEM**

In this project we are giving brief review on various machine learning algorithms such as SVM, Random Forest, Decision Tree, CNN, KNN, MLP and many more to predict spam emails. We should done experiments with above algorithms by using various SPAM datasets such as SPAM ARCHIVE, SPAMBASE, LINGSPAM, PU1 and many more but we are using SPAMBASE dataset to evaluate performance of above algorithms in terms of accuracy, precision and recall.

**Advantages of Proposed System:**

1. Security is more compared to existing systems.
2. Accuracy is more.

**Modules:**

1. Upload SpamBase Dataset

2.Preprocess Dataset

3. Run KNN, Naive Bayes & Multilayer Perceptron Algorithms

4.Run SVM, Decision Tree & AdaBoost Algorithms’

5.Run Random Forest & CNN Algorithm

6. Accuracy Comparison Graph

7.Recall Comparison Graph’

8.Precision Comparison Graph

1. Upload SpamBase Dataset

The selecting and uploading ‘spambase.data’ dataset and then click on ‘Open’ button to load dataset. Then the dataset may loaded.

2.Preprocess Dataset

Preprocessing is the second module in our project. To read all values from dataset and then split data into train and test part where application used 80% dataset for training and 20% dataset for testing.

3. Run KNN, Naive Bayes & Multilayer Perceptron Algorithms

We have to run all 3 algorithms and get there prediction metrics, we got evaluation metrics such as accuracy, recall and precision for all 3 algorithms.

4.Run SVM, Decision Tree & AdaBoost Algorithms’

First we have to run Run SVM, Decision Tree & AdaBoost Algorithms. Then we got metrics for SVM, decision tree and AdaBoost algorithms.

5.Run Random Forest & CNN Algorithm

We should run Random Forest & CNN Algorithm, then we got accuracy for CNN and random forest algorithms.

6. Accuracy Comparison Graph

In graph x-axis represents algorithm name and y-axis represents accuracy of all those algorithms and from above graph we can conclude that MLP neural network give better prediction accuracy compare to all other algorithms.

7.Recall Comparison Graph’

In graph x-axis represents algorithm name and y-axis represents Recall values of all those algorithms .

8.Precision Comparison Graph

In graph x-axis represents algorithm name and y-axis represents Precision values of all those algorithms .

**SYSTEM REQUIREMENTS**

**HARDWARE REQUIREMENTS:**

# Processor - Pentium –IV

* Speed - 1.1 Ghz
* RAM - 256 MB(min)
* Hard Disk - 20 GB
* Key Board - Standard Windows Keyboard
* Mouse - Two or Three Button Mouse
* Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* Operating System - Windows7/8
* Programming Language - Python