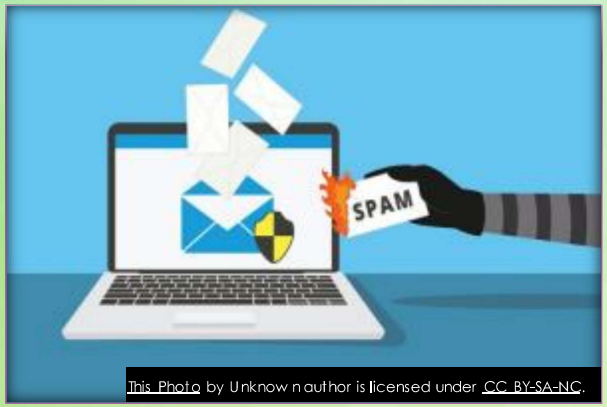
MIT COLLEGE OF ARTS AND SCIENCE FOR WOMEN

SPAM MAIL PREDICTION



INTRODUCTION

- Email is an effective, faster and cheaper way of communication. It is expected that the total number of worldwide email accounts is increased from 3.3 billion email accounts in 2012 to over 4.3 billion by the end of the year 2016.
- Spam is an unwanted , junk, unsolicited bulk mails, used to spreading virus, trojans, malicius code, advertisement or to gain profit on negligible cost.
- Ham is a legitimate, wanted, solicited mails.

E-MAIL SPAMMING

Email spamming is increasing day by day because of effective, fast and cheap way of exchanging information with each other. According to the investigation, user receives.

Spam mails>ham mails

About 120 billion of spam mails are sent pe day and the cost of sending is approximately zero.

Spam is a major problem that attacks the exisence mails. so, it is very important to distinguish ham emails from spam emails, many methods have been proposed for classification of emails as spam or ham emails.

CLASSIFICATION

- Claassification is a predictive modeling. classification consist of assigning a class label to a set of unclassified cases as in

STEPS OF CLASSIFICATION

1. **MODEL CONSTRUCTION:**

Each tuple/sample is assumed to belong to a predefined class, as determined by the class attribute.

The set of tuples used for model construction is traning set.

2. **MODEL USAGE:**

Estimate accuracy of the model.

If the accuracy is acceptable, use the model to classify new data.

SCOP OF THE PROJECT

- It provides senditivity to the client and adapts well to the future spam techniques.
- It considers a complete instead of single words with respect to its organization.
- It increases security and control.
- It reduces IT administration costs.
- It also reduce network resource costs.

NAÏVE BAYES CLASSIFIER

- The naïve bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions.
- The independence assumptions often do not have an impact on reality. Therefore they are considered as naïve.
- It is very useful to classify the e-mails properly.
- The precision and recall of this method is known to be very effective.

SVM

"Support Vector Machine"(svm) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly, used in classification problems.

Svm classifier mostly used in addressing multi-classification problems (having more than 2 target classes to predict).

CONCLUSIO N

We are able to classify th emails as spam or non-spam. Wtith high number of emails lots if people using the system it will be difficult to handle all possible mails as our project deals with only limited amount of corpus.

Thank you