**A study on Email spam detection using**

**supervised learning algorithms**

P Deepika1\*, D Siva Krishna

Under Graduation of computer Science

Engineering Institute of Technology,

Assistant Professor Mr. D. Siva Krishna,

Department of CSE, GMR Institute of Technology

**Abstract.** Social networking helps to collaborate, contribute and connect. Social network is prevalent that all people across the globe are visible to anyone and anywhere. Increase in growth of interest in various social network platforms lead to the huge number of interactions between the users to users or users to websites. Among all, most of the business and general communication agents are working through email because of its cost effectiveness as sending an email is easy and cheap.This leads various attacks like Spamming, Phishing email, Spear Phishing, Link Manipulation, Fake Websites,CEO fraud, Content injection and many more. Therefore, detecting of these spam mails that were fraud is of the most important. Spam detection methods can be broadly divided into expert-based and machine learning based detection methods. In this study, it aims the detection of spam emails using machine learning techniques which improves a way in social network analysis. In this work, spam detection includes different Machine Learning Techniques such as supervised and unsupervised learning. In supervised learning ,K-Nearest Neighbor (KNN), Naive Bayes, Decision Trees, Support Vector Machine (SVM) are performed. Finally, at the end of study, the comparison of different spam email detection techniques will be presented and demonstrates the overall performances of all algorithms regarding accuracy rate.

**Keywords:** *Spamming, Supervised, Classification, Naïve Bayes,Decision tree,Accuracy rate.*

# Introduction

Over the past few years, the Internet has been leaping forward, and the intelligent terminals have been progressively popularized and in this form, Machine learning models have been utilized for multiple purposes in the field of computer science from resolving a network traffic issue to detecting a malware. Coming to digital communications, Email is an primary medium throughout the world. Every personal, social and business communication needs Email and malware is spamming. Email spamming is generally defined as the act of dispersing messages that are unsolicited sent in bulk, using the medium of email. On the other side, emails that are communicated for genuine, lawful and authorised and legitimate purposes are defined as Ham . Spammers use the act of spamming for not only marketing purposes, but also to achieve more malicious goals such as reputational damage and financial disruption, both in institutional and personal front.

# METHODOLOGY:

* 1. **Supervised Machine Learning**

Supervised machine learning is a component of machine learning. Supervised learning is defined as machines trained using well-labeled "well-labeled" training data, based on that labeled data, machines predict output. Labeled data means that some input data is already marked with the correct output.Thus, supervised reading is the process of providing input data as well as positive output data from a machine learning model. The purpose of a supervised learning algorithm is to find a map function to map to input variables (x) and output variables (y).

*EXAMPLES :* **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized. Supervised learning can be separated into two types: 1)Classification 2)Regression

So, Spam detection comes under the classification that classifies the email as Spam or Ham. The algorithms used are : 1) K-Nearest Neighbor (KNN), 2)Naive Bayes, 3)Decision Tree 4)Support Vector Machine (SVM)

## K-Nearest Neighbour Algorithm(KNN):

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.K-NN is a **non-parametric algorithm,** which means it does not make any assumption on underlying data.It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

*Working :*

**Step-1:** Select the number K of the neighbors

**Step-2:** Calculate the Euclidean distance of **K number of neighbors**

**Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.

**Step-4:** Among these k neighbors, count the number of the data points in each category.

**Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.

**Step-6:** Our model is ready.

There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.Large values for K are good, but it may find some difficulties.

* 1. **Naïve Bayes :**

Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.It is mainly used in text classification that includes a high-dimensional training dataset.Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.**It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.

*EXAMPLES :* **spam filtration, Sentimental analysis, and classifying articles**.

Each feature individually contributes to identify that it is an apple without depending on each other.

The formula for Bayes' theorem is given as:

**P(B/A)=P(A/B)\*P(B)/P(A)**

**Where,**

**P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.

**P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true.

**P(A) is Prior Probability**: Probability of hypothesis before observing the evidence.

**P(B) is Marginal Probability**: Probability of Evidence.

* 1. **Decision Tree :**

Decision Tree is a**Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.The decisions or the test are performed on the basis of features of the given dataset.It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree :

**Figure 1. General structure of a Decision Tree**

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.The logic behind the decision tree can be easily understood because it shows a tree-like structure.

*Working :*In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

**Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.

**Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**

**Step-3:** Divide the S into subsets that contains possible values for the best attributes.

**Step-4:** Generate the decision tree node, which contains the best attribute.

**Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

## *Attribute Selection Measures :*

## While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as ****Attribute selection measure or ASM.****By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are: 1)**Information Gain 2)Gini Index**

*1)Information Gain :*Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.

Information Gain= Entropy(S)- [(Weighted Avg) \*Entropy(each feature)

***Entropy :*** Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

***Where,***

**S= Total number of samples, P(yes)= probability of yes, P(no)= probability of no**

### *2)Gini Index :* Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.

Gini Index= 1- ∑jPj2

## *Pruning:* Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of tree **pruning**technology used: **1)Cost Complexity Pruning 2)Reduced Error Pruning.**

* 1. **Support Vector Machine(SVM) :**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

***Hyperplane:*** There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

***Support Vectors:***The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

# RESULTS AND DISCUSSIONS

A framework with the combination of the experts and machine learning techniques collaboratively performed well in detecting spam on social networks.It is difficult to detect spam using only machine learning because of various problems encountered in reality. When only experts are involved in spam detection, it leads to the more time-consuming or costly expenses that can be problematic. The study results in several different observations especially in the realm of Machine Learning based proposition. It is observed that better consistency in the performance of the model is through supervised approaches.SVM and Naïve Bayes are performed well comparing to Decision Tree and KNN. Support Vector Machines (SVM) performed with the ‘Test Accuracy’ of 97. 44% and Naive Bayes (NB) with 94. 57%.

**Figure 2.Pie Chart of Different Machine Learning Algorithms**

# CONCLUSION

In a security point of view, classification of emails as a spam and ham has most important for the users. Machine Learning plays a key role in this classification process for detecting the Spam Mail. All the classification techniques have to be trained first in separating spam emails from other emails before they are actually used. A data set called training set is used to train these techniques. Thousands of samples are used in these training set to make the classifier able to separate the spam mail. But even after this much work spam mail still persists. They persist because every day a new kind of spam mail is introduced. Thus, even if we get old spam mail sorted and marked, new one keep coming in. One of the solutions is to make the training set up-to-date by gathering information about the new kind of spam mail. The fastest way to do that is to make the user report the spam mail as soon as they encounter it and contribute to the global training set because it will take time if the service provided has to monitor each and every mailbox on their own to search for any new spam mail. This algorithm is expected to raise the efficiency of other techniques by some margin depending on the technique. If it is successful in doing so we will have the spam mail dealt with before it reaches our mailbox. This will also save our time and inbox will be less crowded thus making it easier to find useful emails.

# REFERENCES

[1] J. Choi and C. Jeon, "Cost-Based Heterogeneous Learning Framework for Real-Time Spam Detection in Social Networks With Expert Decisions," in IEEE Access, vol. 9, pp. 103573-103587, 2021, doi: 10.1109/ACCESS.2021.3098799.

[2] A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti and M. Alazab, "A Comprehensive Survey for Intelligent Spam Email Detection," in IEEE Access, vol. 7, pp. 168261-168295, 2019, doi: 10.1109/ACCESS.2019.2954791.

[3] S. Gibson, B. Issac, L. Zhang and S. M. Jacob, "Detecting Spam Email With Machine Learning Optimized With Bio-Inspired Metaheuristic Algorithms," in IEEE Access, vol. 8, pp. 187914-187932, 2020, doi: 10.1109/ACCESS.2020.3030751.

[4] Z. Zhang, R. Hou and J. Yang, "Detection of Social Network Spam Based on Improved Extreme Learning Machine," in IEEE Access, vol. 8, pp. 112003-112014, 2020, doi: 10.1109/ACCESS.2020.3002940.

[5] T. Xia, "A Constant Time Complexity Spam Detection Algorithm for Boosting Throughput on Rule-Based Filtering Systems," in IEEE Access, vol. 8, pp. 82653-82661, 2020, doi: 10.1109/ACCESS.2020.2991328.

[6] Rakesh Nayak and Salim {Amirali Jiwani} and B. Rajitha, " Spam email detection using machine learning algorithm",in Sciencedirect,Materials Today: Proceedings,ISSN 2214-7853,https://doi.org/10.1016/j.matpr.2021.03.147,2021.