GMR Institute of Technology
An Autonomous Institute Affiliated to JNTUK, Kakinada

GMRIT
Training Tomorrow's
Engineers Today

A study on Email Spam Detection using Supervised Learning Algorithms

A Term paper report submitted in partial fulfillment of the requirement for the Award of degree

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted
By

Potnuru Deepika
19341A05D5

Under the esteemed guidance of

Mr.D.Siva Krishna
Assistant Professor,
Dept. of Computer Science &Engineering

Department of Computer Science and Engineering

GMR INSTITUTE OF TECHNOLOGY
(An Autonomous institute, affiliated to J.N.T.University kakinada)
NAAC “A” Graded, NBA Accredited, ISO 9001:2008 Certified Institution
G.M.R. Nagar, Rajam-532127, A.P
2021-22

GMR Institute of Technology
An Autonomous Institute Affiliated to JNTUK, Kakinada

GMRIT
Training Tomorrow's
Engineers Today

GMR INSTITUTE OF TECHNOLOGY

(An Autonomous institute, affiliated to J.N.T. University Kakinada)
NAAC "A" Graded, NBA Accredited, ISO 9001:2008 Certified Institution
G.M.R. Nagar, Rajam-532127, A.P

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that term paper report titled **"A STUDY ON EMAIL SPAM DETECTION USING SUPERVISED LEARNING ALGORITHMS"** *submitted by* **POTNURU DEEPIKA** *bearing* **19341A05D5** *has been carried out in partial fulfillment forthe award of* **B.Tech** *degree in the discipline of* ***Computer Science &Engineering*** *to* **JNTUK** *is a record of bonafide work carried out under our guidance and supervision.*

The report embodied in this paper has not submitted to any other university or institution for the award of any degree or diploma.

Signature of the Supervisor Signature of the H.O.D

Mr.D.Siva Krishna
Assistant Professor
Department of CSE
GMRIT, Rajam.

Dr.A V Ramana,
Professor and Head
Department of CSE
GMRIT, Rajam.

ACKNOWLEDGEMENT

It gives me an immense pleasure to express deep sense of gratitude to my guide, **Mr.D.Siva Krishna, Assistant Professor**, Department of Computer Science and Engineering of whole hearted and invaluable guidance throughout the report. Without his sustained and sincere effort, this report would not have taken this shape. He encouraged and helped me to overcome various difficulties that I have faced at various stages of my report.

I would like to sincerely thank **Dr. A V Ramana,** Professor & HOD, Department of Computer Science and Engineering, for providing all the necessary facilities that led to the successful completion of my report.

I take privilege to thank our Principal **Dr. C.L.V.R.S.V.Prasad**, who has made the atmosphere so easy to work. I shall always be indebted to them.

I would like to thank all the faculty members of the Department of Computer Science and Engineering for their direct or indirect support and also all the lab technicians for their valuable suggestions and providing excellent opportunities in completion of this report.

P.Deepika
19341A05D5

INDEX

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.With increased use of internet, numbers of email users are increasing day by day. This increasing use of email has created problems caused by unsolicited bulk email messages commonly referred to as Spam. Malware is spamming.

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. There are many effects of Spam. It fills our inbox with number of ridiculous emails. Spammers steals useful information like details on you contact list and alters your search results on any computer program. Spam is a huge waste of everybody’s time and can quickly become very frustrating if you receive large number of spam emails .Identifying these spammers and the spam email is a laborious task.

REASON FOR SPAMMING :

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. Besides advertising, these may contain links to phishing or malware hosting websites found out to steal confidential information.

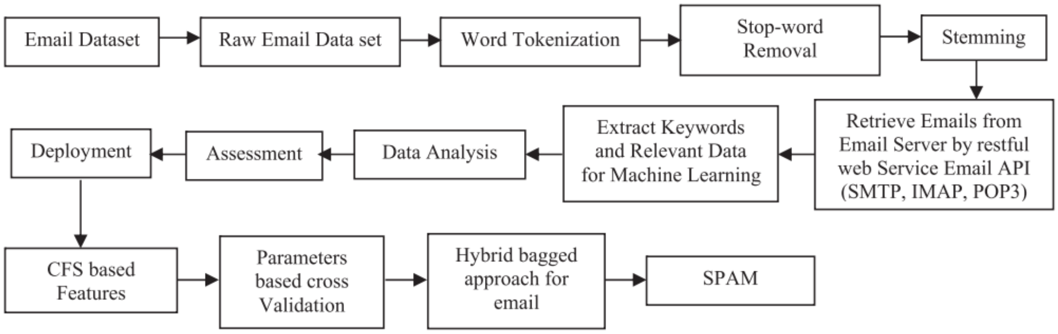
INTRODUCTION TO MACHINE LEARNING :

Machine learning is a branch of artificial intelligence (AI) and computer science which mainly focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

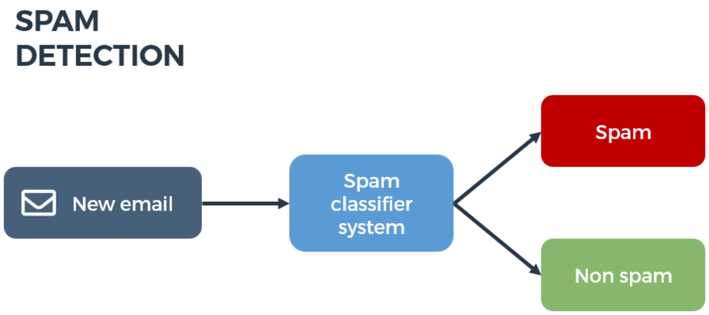
Many tools and techniques are offered by companies in order to detect spam emails in a network. Organisations have set up filtering mechanisms to detect unsolicited emails by setting up rules and configuring the firewall settings.To solve this problem, Machine learning plays a major in doing classification of spam and ham.

STEPS IN MACHINE LEARNING :

The machine learning algorithmic method’s affectivity inspires the proposed spam email detection system. Firstly collection of email data is accomplished in the spam mail detection system. The nature of unconstructed and raw is having in those collected email data. It is required to pre-process the email data for the optimal computations in addition to achieve the accurate results.


Email Dataset
Raw Email Data set
Word Tokenization
Stop-word Removal
Stemming
Retrieve Emails from Email Server by restful web Service Email API (SMTP, IMAP, POP3)
Extract Keywords and Relevant Data for Machine Learning
Data Analysis
Assessment
Deployment
CFS based Features
Parameters based cross Validation
Hybrid bagged approach for email
SPAM

DIFFERENT ALGORITHMS IN SPAM DETECTION IN MACHINE LEARNING :



Machine learning has best algorithms in detecting the spam or ham emails. It has different supervised and unsupervised learning algorithms. Some of them are namely, Random Forest, Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Machine(SVM), K-Nearest Neighbour (K-NN) .By using machine learning algorithms,vspam and ham can be detected with greater accuracy.

LITERATURE SURVEY

[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.

Detection of spam email by human intervention is difficult as it is time consuming and cost effective. Depending on machine learning techniques also have some constraints like frequent updating of keywords and blacklist, mismatch of ham and spam. So, A sophisticated framework is designed in which experts and machine learning algorithms collaborate worked together to detect spam filed effectively. The paper have shown by experiments that it is difficult to detect spam using only machine learning because of various problems encountered in reality. So,author proposed a collaborative system including both machine learning and experts for spam detection.

Detection techniques using machine learning have the advantage of being automated, even though the features and techniques used are different. It has supervised learning, which is static and requires experts to annotate each set of data. It is ironic that machine learning techniques that seek to minimize human intervention costs to prepare datasets. Techniques using machine learning have various problems in addition to the fundamental problem of having to be preceded by human effort. Detection of spam methods in which some or many experts participated and detect it manually. These methods have exceptionally high accuracy because experts directly participate and make decisions. However, it takes much time to gather the opinions of experts.

The author has used tree-based J48, random forest (RF), rule-based PART, Naïve Bayes network algorithms in Machine Learning. The author has two datasets spam datasets. First dataset is combination of normal and spam data in the ratio of 1:1. The second dataset is composed of normal and spam data in a 95:5 ratio. Finally, the collaborate framework had performed well in detecting the spam among all data.

[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.

There are number of attacks that are being constantly bombarded on users worldwide, such as email spoofing phishing, variants of phishing attacks within spam emails. So,the author of this paper focused about Artificial Intelligence (AI) and Machine Learning (ML) methods for intelligent spam email detection. The authors has considered 4 parts in the email's structure for analysis.

There are of : 1)Headers Provide Routing Information

2)SMTP(Simple Mail Transfer Protocol) Envelope

3) First part of SMTP Data

4) Second part of SMTP Data.

The first part contains of mail transfer agents (MTA) that provide information like email and IP address of each sender and recipient of where the email originated and final destination reached. The Second part tells, the SMTP Envelope that containing mail exchanger's identification, originating source and destination domains\users. The third part in detections is First part of SMTP Data that containing information like from, to, date, subject that appears in most emails. Final part in the email's structure for analysis i.e. Second part of SMTP Data containing email body including text content, and attachment. The author performed Supervised algorithms that are Naïve Bayes, Decision Tree, Support Vector Machine(SVM), Logistic Regression on four different spam datasets. The author has highlighted that certain algorithms such as SVM and Naïve Bayes are in high demand. The paper gave conclusion that single-algorithm anti-spam systems are quite common and the potentiality of research into hybrid and multi-algorithm systems is quite promising.

[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.

Emails are used regularly by many people around the world for communication and for socializing. This gained to unauthorized access to the device by tracking the user into clicking the spam link within the spam email and leads to various malicious activities. Many tools and techniques were used in order to detect spam emails in a network. Machine Learning is the best among all the techniques. This paper aimed to detect the spam emails with machine learning algorithms that are optimized with bio-inspired methods. The authors have conducted experiments with six different machine learning algorithms namely Naïve Bayes (NB) classification, K-Nearest Neighbour (K-NN), Artificial Neural Network (ANN), Support Vector Machine (SVM), Artificial Immune System and Rough Sets. Tokenisation was performed during detection of spam mails. Each algorithm has two stage, one is Training and the other one is Filtering. The algorithm consisted of four steps : 1) Email Pre-Processing

2) Description of the feature

3) Spam Classification and

4) Performance Evaluation.

Various ratios of training and testing were performed to get the best ratio among 60:40, 70:30, 75:25, 80:20 and 80:20 was chosen as it got good F1-score, Precision and Recall in comparison to Accuracy. Among all the six machine learning algorithms, authors concluded that the Naïve Bayes provided the highest accuracy, precision and recall. Later, a hybrid system were constructed between two machine learning algorithms those are of SVM & NB. The method is to apply the SVM algorithm and generate the hyperplane between the given dimensions and reduce the training set by eliminating datapoints. Then after, this set will then be implemented with NB algorithm to predict the probability of the outcome. This system was successfully implemented and there was an increase in accuracy when compared to NB and SVM on their own. This paper experimented on two types of spam email dataset, alphabetic-based and numeric-based datasets.

[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.

Online Social Networks (OSN) turns out to be a critical channel for people to acquire information, disseminate information, and make friends and get entertained. This paper took example of Twitter and worked on the twitter dataset to identify spam tweets. The author has characterized the details like

a)Feature Selection (content-based and user profile-based)

b) In algorithm selection

This paper has focused primarily on the use supervised machine learning algorithms to deal with spam detection in social networks. As spam detection is under the classification category, the researchers have designed numerical form characteristics to identify spam users. The supervised machine learning algorithm can be split into a single classification algorithm and an integrated classification algorithm that are of Support Vector Machine (SVM), Naive Bayesian (NB),K-nearest Neighbor (KNN),Decision Tree (DT) and Random Forest (RF). In this paper, the author focused on spam in twitter in which the detection done on user attribute, content, activity and relationship of each accounts. In the features used by the author are of age of account, number of tweets, number of retweets, number of hashtags, URLs, character, digits, spam words, time between replies, Number of followers, followings and favorites. It has considered both balanced and unbalanced datasets and calculated the performance by supervised algorithms. In this paper,the author has took two datasets to compare the experimental results. The first dataset is the public dataset namely the Aponador dataset.Such dataset was collected with Brazil’s famous location-based social network and covers both normal users and spam users, in which each record contains 59 characteristic and 2 classifications. The second dataset is harvested by this study using the Twitter API and Twitter4J library, which cover 43 million tweets posted by around 16 million accounts that contains daily popular trends.

[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.

Internet had connected each individual together by computers or mobile devices. Along with it, the scale of data is overwhelmingly increased as well especially after the wide use of social networks, personal communication tools, emails and short messages (SMS). This easy-communication circumstance also encouraged the numerous emerge of spams. Such kind of activities turned into one of the most profitable businesses for spammers. This paper addresses the challenging throughput issue and proposes a constant time complexity rule-based spam detection algorithm.The author stated that machine learning approaches usually require a beginning training for spam filter and training again if rules are updated. Then, the filter may be required to restart to load the updated models. The content filtering systems utilize statistical machine learning approaches. Many models have been applied to obtain better spam detection results. Some of the them are Support Vector Machine (SVM), Bayesian methods and Decision Tree. SVM creates a multiclass and trains a decision equation from a high-dimensional feature space, which leads to get high accuracy. Bayes, are regarded as the effective and important machine learning algorithms in information retrieval. Bayesian algorithm proposed to defeat spammers.

Machine learning approaches usually require a beginning training for spam filter and training again if rules are updated. Then, the filter may be required to restart to load the updated models. The author has done experiment to validate the constant time complexity of the proposed spam detection algorithm.A constant time complexity spam detection algorithm was completed to boost throughout on rule-based filtering systems with overall time complexity for spam detection is O(1). Sequential Matching is an additional feature of the method presented in this paper. The author has concluded that the speed of the spam detection algorithm presented in this paper is independent of rule size and rule term vocabulary.

[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

In the internet, this spam became a immense adversity. Generally, the waste of message speed, time and storage is known as a Spam. The Machine learning approach is the most effective and efficient one which uses a training dataset. Training datasets are the samples that take a set of pre-classified emails. This includes the algorithms like support vector machines, K-nearest neighbor, Naïve Bayes, Random Forests. Experiments can be performed by using the email dataset collected from the kaggle website library.After dataset collection through kaggle, author has specified the preprocessing on dataset such as tokenization, stop word removal. The researchers conducted experiments on email dataset by performing the various algorithms .They are of K-Nearest Neighbor (K-NN), Support Vector Machine (SVM) and Naive Bayes classification (NB). The author has used four steps in the algorithms.

They are: 1) pre-processing of email dataset

2) feature description

3) classification of spam and

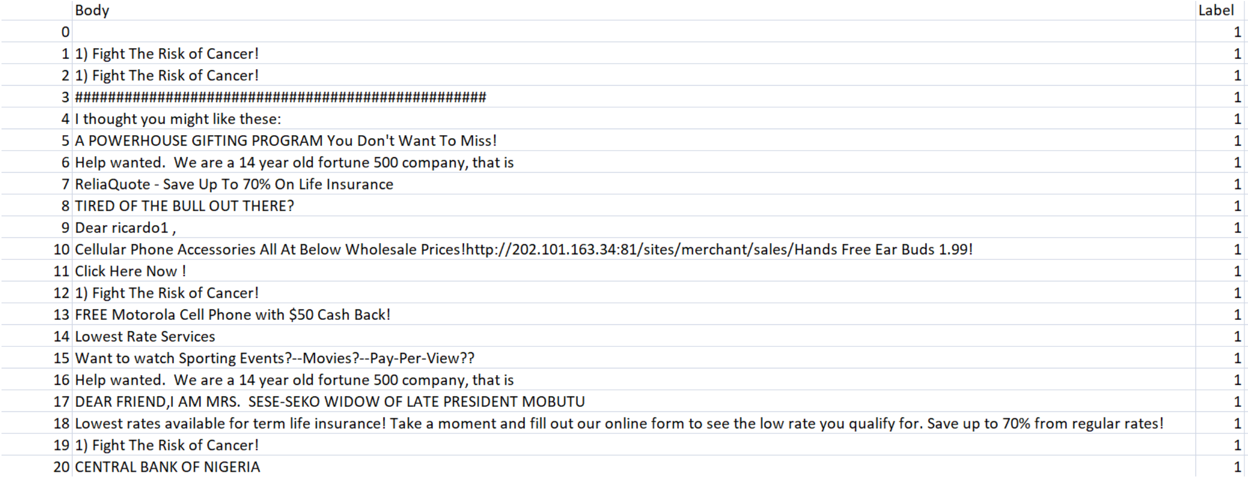
4) evaluation of system performance.

The author of paper concluded that the utmost precision, recall and accuracy were offered by using the Naive Bayes algorithm.

DATASETS

Dataset 1:-

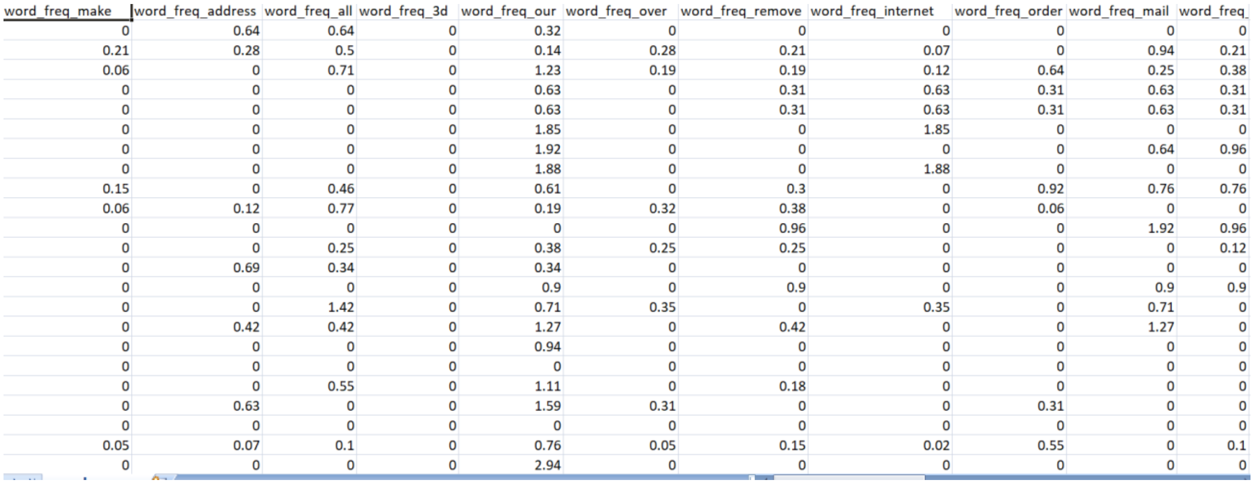
This dataset contains of 3 columns i.e. S.No, Body of the Email, Label.



	Body	Label
0		1
1	1) Fight The Risk of Cancer!	1
2	1) Fight The Risk of Cancer!	1
3	##	1
4	I thought you might like these:	1
5	A POWERHOUSE GIFTING PROGRAM You Don't Want To Miss!	1
6	Help wanted. We are a 14 year old fortune 500 company, that is	1
7	ReliaQuote - Save Up To 70% On Life Insurance	1
8	TIRED OF THE BULL OUT THERE?	1
9	Dear ricardo1 ,	1
10	Cellular Phone Accessories All At Below Wholesale Prices!http://202.101.163.34:81/sites/merchant/sales/Hands Free Ear Buds 1.99!	1
11	Click Here Now !	1
12	1) Fight The Risk of Cancer!	1
13	FREE Motorola Cell Phone with $50 Cash Back!	1
14	Lowest Rate Services	1
15	Want to watch Sporting Events?--Movies?--Pay-Per-View??	1
16	Help wanted. We are a 14 year old fortune 500 company, that is	1
17	DEAR FRIEND,I AM MRS. SESE-SEKO WIDOW OF LATE PRESIDENT MOBUTU	1
18	Lowest rates available for term life insurance! Take a moment and fill out our online form to see the low rate you qualify for. Save up to 70% from regular rates!	1
19	1) Fight The Risk of Cancer!	1
20	CENTRAL BANK OF NIGERIA	1

Dataset 2:-

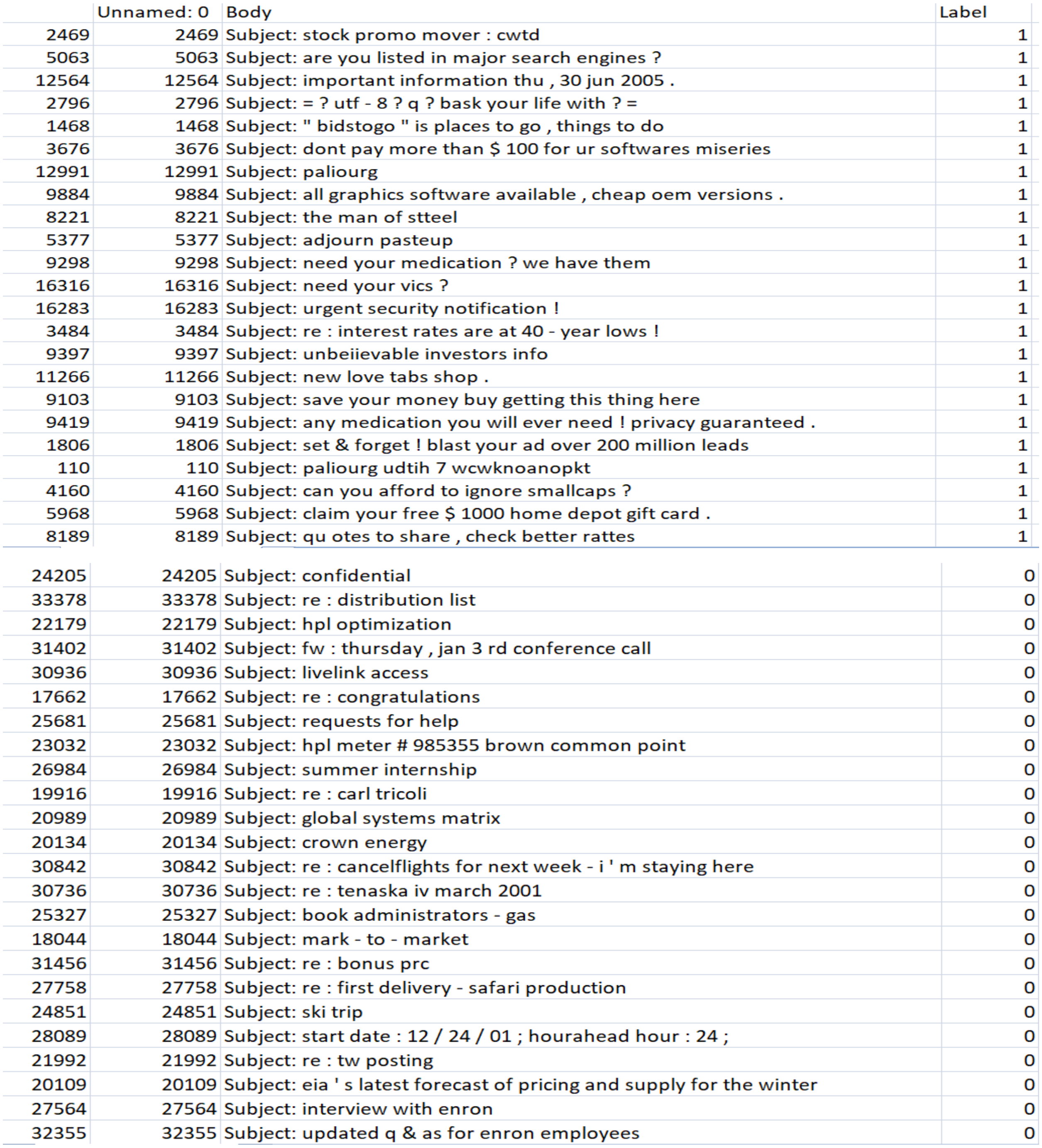
This dataset contains of 26 columns of frequent words along with their frequency.



word_freq_make	word_freq_address	word_freq_all	word_freq_3d	word_freq_our	word_freq_over	word_freq_remove	word_freq_internet	word_freq_order	word_freq_mail	word_freq_
0	0.64	0.64	0	0.32	0	0	0	0	0	0
0.21	0.28	0.5	0	0.14	0.28	0.21	0.07	0	0.94	0.21
0.06	0	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38
0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31
0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31
0	0	0	0	1.85	0	0	1.85	0	0	0
0	0	0	0	1.92	0	0	0	0	0.64	0.96
0	0	0	0	1.88	0	0	1.88	0	0	0
0.15	0	0.46	0	0.61	0	0.3	0	0.92	0.76	0.76
0.06	0.12	0.77	0	0.19	0.32	0.38	0	0.06	0	0
0	0	0	0	0	0	0.96	0	0	1.92	0.96
0	0	0.25	0	0.38	0.25	0.25	0	0	0	0.12
0	0.69	0.34	0	0.34	0	0	0	0	0	0
0	0	0	0	0.9	0	0.9	0	0	0.9	0.9
0	0	1.42	0	0.71	0.35	0	0.35	0	0.71	0
0	0.42	0.42	0	1.27	0	0.42	0	0	1.27	0
0	0	0	0	0.94	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0.55	0	1.11	0	0.18	0	0	0	0
0	0.63	0	0	1.59	0.31	0	0	0.31	0	0
0	0	0	0	0	0	0	0	0	0	0
0.05	0.07	0.1	0	0.76	0.05	0.15	0.02	0.55	0	0.1
0	0	0	0	2.94	0	0	0	0	0	0

Dataset 3:-

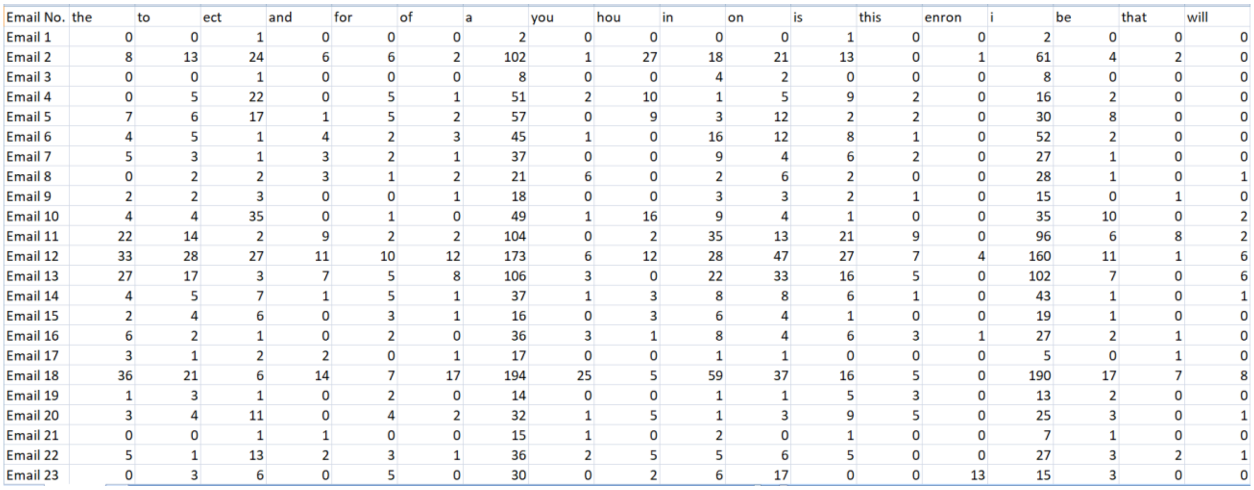
This dataset contains of 4 columns with label as 1 with it is spam and 0 if it is ham mail.



	Unnamed: 0	Body	Label
2469	2469	Subject: stock promo mover : cwtd	1
5063	5063	Subject: are you listed in major search engines ?	1
12564	12564	Subject: important information thu , 30 jun 2005 .	1
2796	2796	Subject: = ? utf - 8 ? q ? bask your life with ? =	1
1468	1468	Subject: " bidstogo " is places to go , things to do	1
3676	3676	Subject: dont pay more than $ 100 for ur softwares miseries	1
12991	12991	Subject: paliourg	1
9884	9884	Subject: all graphics software available , cheap oem versions .	1
8221	8221	Subject: the man of stteel	1
5377	5377	Subject: adjourn pasteup	1
9298	9298	Subject: need your medication ? we have them	1
16316	16316	Subject: need your vics ?	1
16283	16283	Subject: urgent security notification !	1
3484	3484	Subject: re : interest rates are at 40 - year lows !	1
9397	9397	Subject: unbeiievable investors info	1
11266	11266	Subject: new love tabs shop .	1
9103	9103	Subject: save your money buy getting this thing here	1
9419	9419	Subject: any medication you will ever need ! privacy guaranteed .	1
1806	1806	Subject: set & forget ! blast your ad over 200 million leads	1
110	110	Subject: paliourg udtih 7 wcwknoanopkt	1
4160	4160	Subject: can you afford to ignore smallcaps ?	1
5968	5968	Subject: claim your free $ 1000 home depot gift card .	1
8189	8189	Subject: qu otes to share , check better rattes	1
24205	24205	Subject: confidential	0
33378	33378	Subject: re : distribution list	0
22179	22179	Subject: hpl optimization	0
31402	31402	Subject: fw : thursday , jan 3 rd conference call	0
30936	30936	Subject: livelink access	0
17662	17662	Subject: re : congratulations	0
25681	25681	Subject: requests for help	0
23032	23032	Subject: hpl meter # 985355 brown common point	0
26984	26984	Subject: summer internship	0
19916	19916	Subject: re : carl tricoli	0
20989	20989	Subject: global systems matrix	0
20134	20134	Subject: crown energy	0
30842	30842	Subject: re : cancelflights for next week - i ' m staying here	0
30736	30736	Subject: re : tenaska iv march 2001	0
25327	25327	Subject: book administrators - gas	0
18044	18044	Subject: mark - to - market	0
31456	31456	Subject: re : bonus prc	0
27758	27758	Subject: re : first delivery - safari production	0
24851	24851	Subject: ski trip	0
28089	28089	Subject: start date : 12 / 24 / 01 ; hourahead hour : 24 ;	0
21992	21992	Subject: re : tw posting	0
20109	20109	Subject: eia ' s latest forecast of pricing and supply for the winter	0
27564	27564	Subject: interview with enron	0
32355	32355	Subject: updated q & as for enron employees	0

Dataset 4:-

This dataset contains of 3000 columns with frequency of specific words.



Email No.	the	to	ect	and	for	of	a	you	hou	in	on	is	this	enron	i	be	that	will
Email 1	0	0	1	0	0	0	2	0	0	0	0	1	0	0	2	0	0	0
Email 2	8	13	24	6	6	2	102	1	27	18	21	13	0	1	61	4	2	0
Email 3	0	0	1	0	0	0	8	0	0	4	2	0	0	0	8	0	0	0
Email 4	0	5	22	0	5	1	51	2	10	1	5	9	2	0	16	2	0	0
Email 5	7	6	17	1	5	2	57	0	9	3	12	2	2	0	30	8	0	0
Email 6	4	5	1	4	2	3	45	1	0	16	12	8	1	0	52	2	0	0
Email 7	5	3	1	3	2	1	37	0	0	9	4	6	2	0	27	1	0	0
Email 8	0	2	2	3	1	2	21	6	0	2	6	2	0	0	28	1	0	1
Email 9	2	2	3	0	0	1	18	0	0	3	3	2	1	0	15	0	1	0
Email 10	4	4	35	0	1	0	49	1	16	9	4	1	0	0	35	10	0	2
Email 11	22	14	2	9	2	2	104	0	2	35	13	21	9	0	96	6	8	2
Email 12	33	28	27	11	10	12	173	6	12	28	47	27	7	4	160	11	1	6
Email 13	27	17	3	7	5	8	106	3	0	22	33	16	5	0	102	7	0	6
Email 14	4	5	7	1	5	1	37	1	3	8	8	6	1	0	43	1	0	1
Email 15	2	4	6	0	3	1	16	0	3	6	4	1	0	0	19	1	0	0
Email 16	6	2	1	0	2	0	36	3	1	8	4	6	3	1	27	2	1	0
Email 17	3	1	2	2	0	1	17	0	0	1	1	0	0	0	5	0	1	0
Email 18	36	21	6	14	7	17	194	25	5	59	37	16	5	0	190	17	7	8
Email 19	1	3	1	0	2	0	14	0	0	1	1	5	3	0	13	2	0	0
Email 20	3	4	11	0	4	2	32	1	5	1	3	9	5	0	25	3	0	1
Email 21	0	0	1	1	0	0	15	1	0	2	0	1	0	0	7	1	0	0
Email 22	5	1	13	2	3	1	36	2	5	5	6	5	0	0	27	3	2	1
Email 23	0	3	6	0	5	0	30	0	2	6	17	0	0	13	15	3	0	0

Table 1: Details of datasets

Dataset number	No. of rows	No.of columns
Dataset 1	22415	3
Dataset 2	4601	58
Dataset 3	10766	4
Dataset 4	5172	3000

METHODOLOGY

SUPERVISED MACHINE LEARNING :

Supervised machine learning is a subcategory of machine learning. Supervised learning is defined as machines that are trained using well "labelled" training data, on basis of that labeled data, machines predict the output. The labelled data means some input data is already tagged with the correct output.So,Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(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)

1)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.

F**itting K-NN classifier to the Training data:**

```
from sklearn.neighbors import KNeighborsClassifier

classifier= KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2)

classifier.fit(x_train, y_train)
```

To fit the K-NN classifier to the training data. To do this we will import the KNeighborsClassifier class of Sklearn Neighbors library. After importing the class, we will create the **Classifier** object of the class. The Parameter of this class will be

- n_neighbors**:** To define the required neighbors of the algorithm. Usually, it takes 5.
- metric='minkowski'**:** This is the default parameter and it decides the distance between the points.
- p=2: It is equivalent to the standard Euclidean metric.

2)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.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem : Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.The formula for Bayes' theorem is given as:

$$\mathbf{P(B/A)=P(A/B)*P(B)/P(A)}$$

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.

Fitting Naive Bayes to the Training Set:

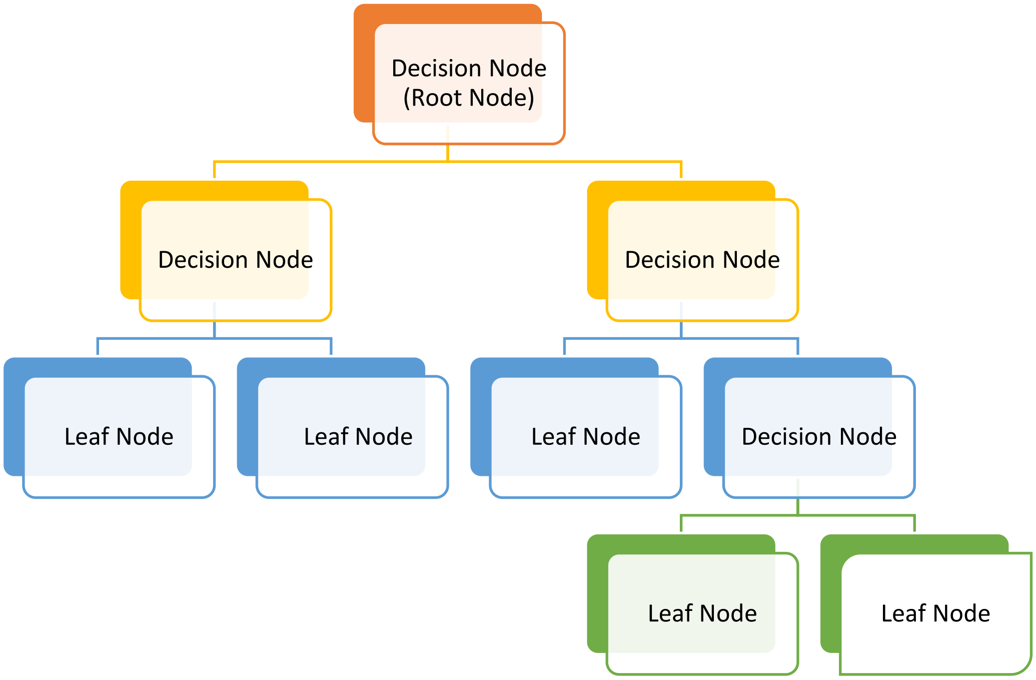
```
from sklearn.naive_bayes import GaussianNB

classifier = GaussianNB()

classifier.fit(x_train, y_train)
```

In the above code, we have used the GaussianNB classifier to fit it to the training dataset. We can also use other classifiers as per our requirement.

3)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 :



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.

Fitting a Decision-Tree algorithm to the Training set :

```
From sklearn.tree import DecisionTreeClassifier

classifier= DecisionTreeClassifier(criterion='entropy', random_state=0)

classifier.fit(x_train, y_train)
```

Import the **DecisionTreeClassifier** class from **sklearn.tree** library."criterion='entropy': Criterion is used to measure the quality of split, which is calculated by information gain given by entropy.random_state=0": For generating the random states.

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- $\sum_j P_j^2$

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

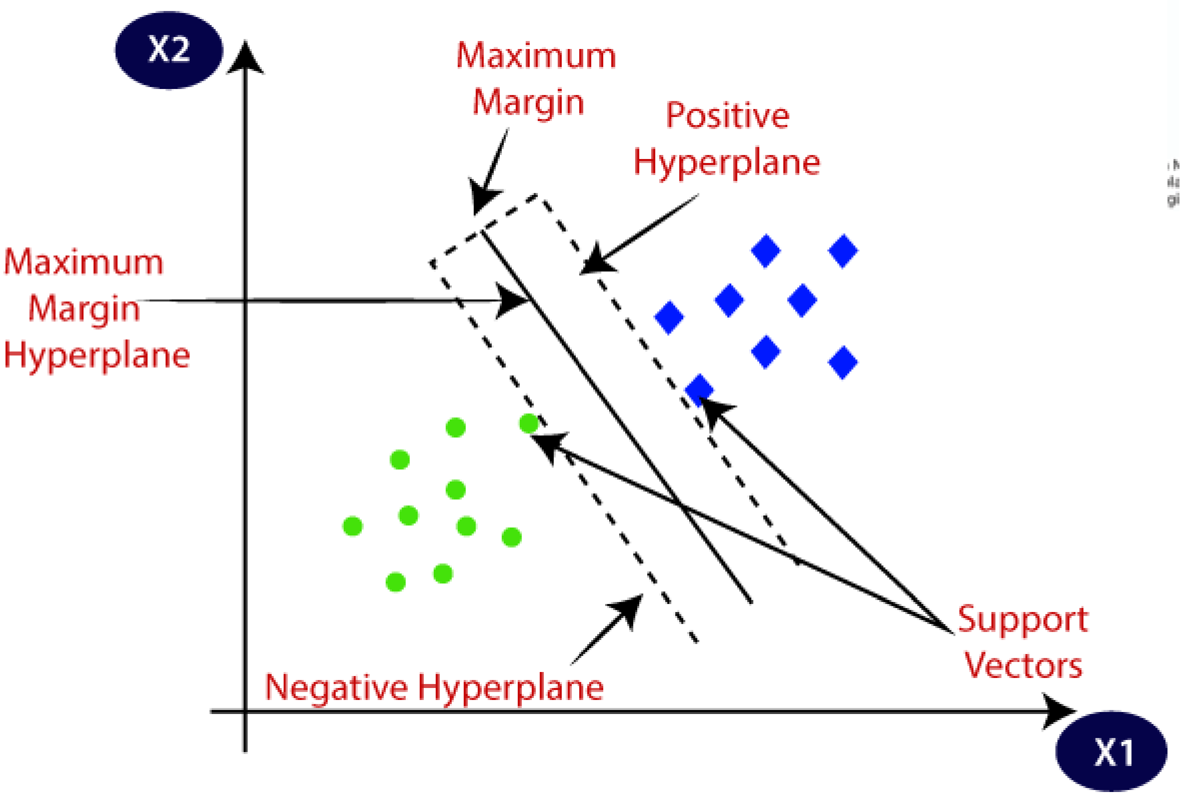
2)Reduced Error Pruning.

4)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. However, primarily, it is used for Classification problems in Machine Learning.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.

Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



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.

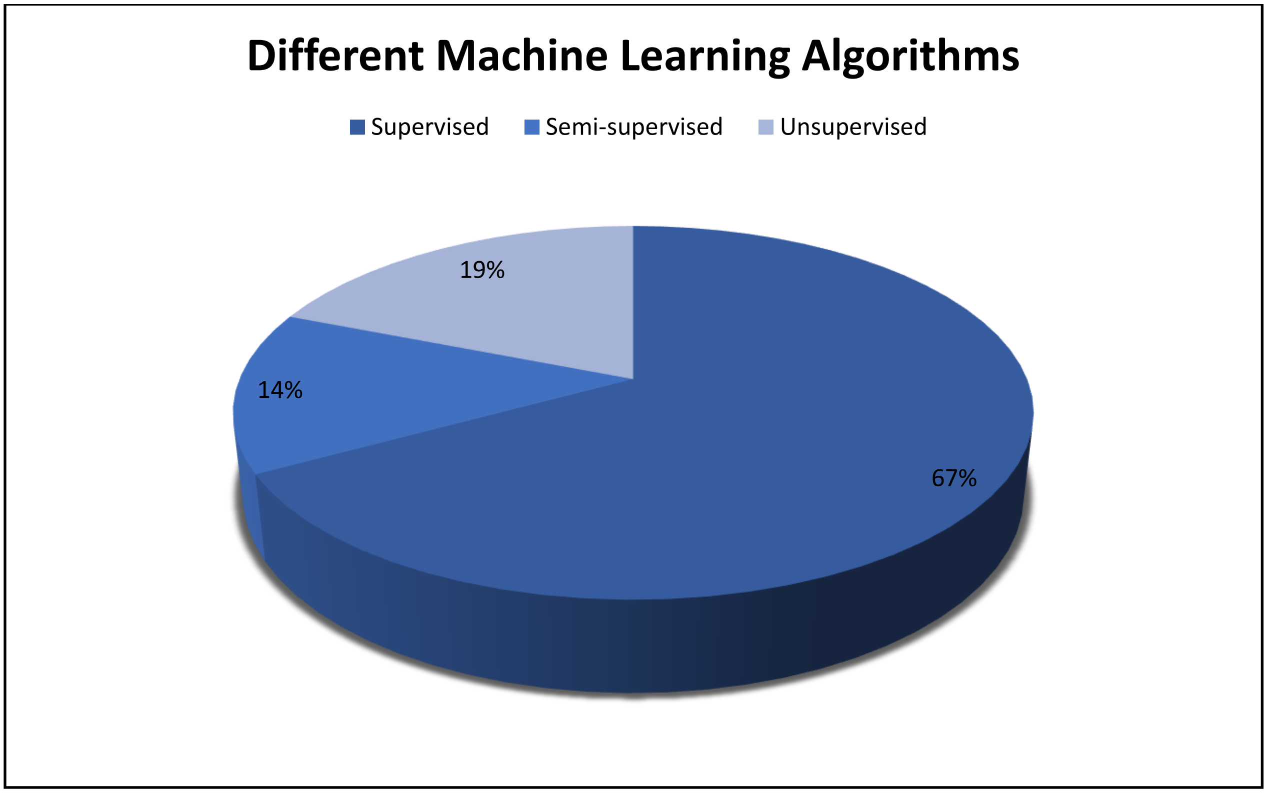
Fitting the SVM classifier to the training set:

```
from sklearn.svm import SVC # "Support vector classifier"

classifier = SVC(kernel='linear', random_state=0)

classifier.fit(x_train, y_train)
```

Import SVC class from Sklearn.svm library In the above code, we have used kernel='linear', as here we are creating SVM for linearly separable data. However, we can change it for non-linear data. And then we fitted the classifier to the training dataset(x_train, y_train).

RESULTS & DISCUSSION

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%.



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