

Thesis: Title: Use of artificial intelligence and data analytics to compare the e-mail spam filtering (YAHOO, Gmail, Outlook) depending upon user experiences

Thesis

Student Name: Shoaib

University Name:

Dated: 25th November 2021

Statement of Authenticity and Wordcount

STATEMENT OF AUTHENTICITY

This project is original research works, and that is originally written by me. This project is not copy paste of other researcher because it shows the originality due to I have put my efforts to find out the extensive machine learning algorithm to meet the corporate challenges. This research is not part of any profit organization and stakeholders. This project does not harm any human being during research work. This research is help out future researcher. My supervisor helps me lot to develop this essential research work.

WORDCOUNT: 12392

Thesis: Title: Use of artificial intelligence and data analytics to compare the e-mail spam filtering (YAHOO, Gmail, Outlook) depending upon user experiences

Thesis

Student Name: Shoaib

University Name:

Dated: 25th November 2021

Abstract ID: _____

Use of artificial intelligence and data analytics to compare the e-mail spam filtering (YAHOO, Gmail, Outlook) depending upon user experiences

Abstract

Email spam detection is common problem for corporate business and enterprises domain, so meet these challenges to eradicate the hackers and spammers, various machine learning algorithm has been widely used to minimize the spam emails attack within the organization. Gmail server works efficiently comparatively to Yahoo and outlook, Gmail server uses Google secures API layer to detect the spammers and hackers, which is based on machine learning and deep learning pattern. In this thesis various machine learning algorithm has been developed naïve bays classifier, multinomial naïve bays, logistics regression , support vector machine, streamlit with Naïve bays has been developed to provides the intended interface of email spams detections. Primary data has been collected from Github repository, and secondary data has been collected from peer reviews journals, such as ACM digital library, IEEE XPLORE sites, Google scholars, science direct. Method uses machine learning algorithm, naïve bays, SVM, LR, TFIDF classifier, countvectorizer, Streamlit, multinomial naïve bays. This project is not part of any funding committee and any stakeholders, this project does not harm any human being during research data collection. Future research required to improve Yahoo and outlook emails servers because various negative comments has been seen in research.

Table of Contents

Thesis: Title: Use of artificial intelligence and data analytics to compare the e-mail spam filtering (YAHOO, Gmail, Outlook) depending upon user experiences	1
Thesis	1
STATEMENT OF AUTHENTICITY	2
WORDCOUNT: 12392	2
Abstract	4
Chapter # 1: Introduction	8
Machine Learning & Deep Learning Method	9
Novel Email Spam Detection Method	10
Spam Email Detection Using Machine Learning & Neural Networks	11
Machine learning algorithm on email spam classification	12
Aim & Objective of this Research	13
Chapter # 2: Literature Review	14
Introduction:	14
Related Work	14
Conclusion of Literature Review	25
Method Used:	25
Chapter # 3 Methodology	26
Introduction	26
Research Philosophy	26
Data Collection and interpretation	27
Research Strategy	27
Gmail Spam Filter:	27
Yahoo Mail Spam Filters:	28
Outlook Email Spam Filter:	28
Data Collection Method	29
Primary Data:	29
Secondary Data:	29
Data Analysis Method	29
Naïve bays classifier:	29
SVM support vector machine classifier	30

Natural Language processing using python with Logistics regression spam email classifier	30
Convolutional Neural Network deep learning algorithm.....	31
Email Spam Classification App that build by using the streamlit & python	31
Quantitative Method	31
• Statistical machine learning method	31
• GUI application	31
Results of Research.....	32
Sources.....	33
Ethical Concern.....	34
Chapter # 4: Results & Analysis	35
Email Spam Classification App that build by using the streamlit & python	35
Streamlit python library installation:	35
Dataset information & Python file.....	36
Data Cleaning & Pre-processing.....	37
Machine Learning Library to build spam models.....	38
Training & testing the dataset	38
Building the Multinomial Naïve Bays Classifier Model	39
Results of Streamlit python Spam Detector in Browser:	40
Analysis:	40
Support Vector Machine Classifier.....	40
Python Libraries:.....	41
Data Set information	41
Split dataset in Training & Testing data	41
Extract Features CountVectorizer:.....	41
Problem with Naïve Bays:	42
Results.....	42
Analysis:	43
Naïve Bays Classifier to Detect Spam Emails.....	44
Importing Python Libraries.....	44
Data set information.....	45
Date Preprocessing.....	45
Training & Testing the dataset.....	45

CountVectorizer to convert words into matrix features.....	47
Multinomial Naïve Bays algorithm to classify text and build the model	47
Results.....	47
Analysis.....	48
Machine Learning Logistics Regression to Detect Spam emails.....	48
Data Preprocessing and Importing CSV	49
Data Cleaning & Text Transformation	52
NLTK library to extract spam words and ham words.....	53
TFID-IDF Vectorizer	54
Transform the string text into numeric array	55
Train & test the data validation.....	55
Outcomes & Results	56
Analysis.....	56
Chapter # 5 Discussion & Conclusion	56
Discoveries.....	57
Novelty.....	57
Future Research Challenges.....	59
Conclusion	60
References.....	61
Reflection.....	67
Research work processes	67
Research idea	67
Skills development.....	68
Teacher meeting.....	68
Research experience.....	68

Use of artificial intelligence and data analytics to compare the e-mail spam filtering (YAHOO, Gmail, Outlook) depending upon user experiences

Chapter # 1: Introduction

Email spam is major issues for the enterprise organization because spam mail contains malware information that executes on the operating system of the use and causes problem (Alzahrani & Rawat, 2019,) spammers and hackers sends various spam emails to the user and steal the essential user data and information. This problem causes an issues to resolve the spam and hack email at the enterprise level the machine learning method used to meeting this challenge. Famous internet service providers including email service provider such as Yahoo, Gmail Outlook uses machine learning method to filter and clean the spam email. It only put the spam email separate email folder, it does not able to detect and delete the spam mails automatically. So the enterprise organization using email server to meet these challenges, being the support of network administrator, the deploy machine learning algorithm in the email server because email server does not facility of machine learning method to detect filter and clean spam email and spam messages in the organization. The development of machine learning AI algorithm in email server that helps and support the corporate to eradicate the spammers attack and hacker attack. The short message is most popular in these days because short sms service gain popularity to build up the policy of mobile communication architecture. Various advertising agencies advertised products using the short messages method, send the unwanted information continuously to the user it might be busy internal memory of system and might be able to steal the data and hack the operating system. The spam messages produces irritation, problem and frustration them it also engaged the user for deletion and reading activity, it causes the problem that wasting the user time and user memory of internal operating system, it might be also able to cost the memory.

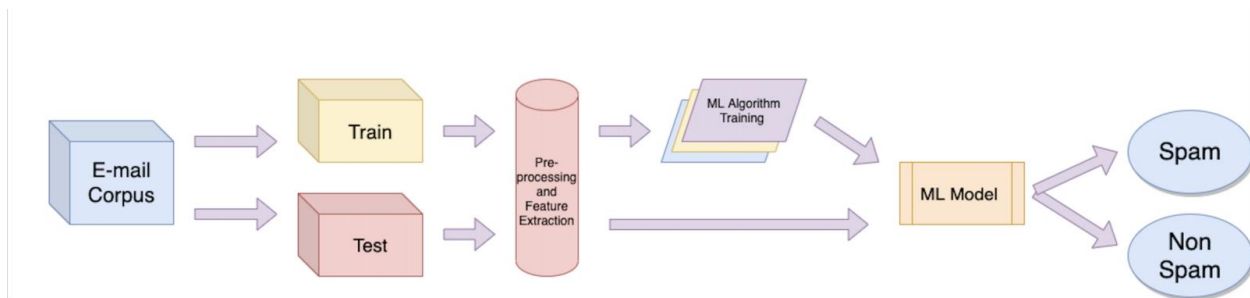


Figure 1: E-mail spam and non-spam filtering

The popular machine learning algorithm method successfully deployed in this thesis to come up the challenge of spam email and spam messages. The essential machine learning algorithm used to detect filter and clean the spam email data such as logistic regression, naïve bays, support vector machine, and neural networks, the method filter the messages and measure the accuracy of most effective machine learning algorithm that deal on spam messaging technique, beside this neural network performs better accuracy due to hidden neural network layer, because neural networks performs well to trained the classifier model and used the method to detect the harm or spam email.

Machine Learning & Deep Learning Method

(Tazmina, et al., 2020) Deep learning works on efficient neural network layer, neural network contains billions of hidden neural layer. The deep learning algorithm works very effective to detect the spam images in the email dataset. The using of convolutional neural network that achieve the highest accuracy by using the image dataset. Beside this machine learning method also works very well by using logistics regression decision tree, random forest tree, naïve bays that achieve the highest accuracy of spam email detection including 98% to 99% of accuracy with F1 score. The uninvited bulk emails which contains malware information and malware dummy link that causes the problematic situation to analyze the text based situation. The image segmentation method widely used to detect the unwanted emails. The text based analysis to detect spam text in large dataset file is also highlighted and deep learning method able to detect the spam text over the large dataset file.

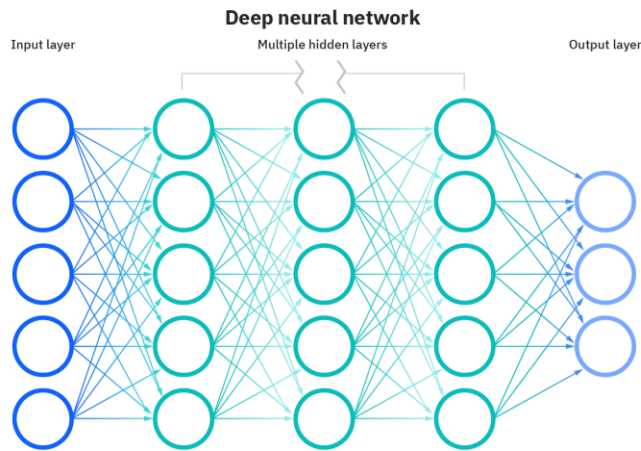


Figure 2: Deep Neural Network

Novel Email Spam Detection Method

(Ezpeleta & Mendizabal, 2020) Unwanted junks spam email is part of every email domain that very difficult to tolerate them at enterprise level. In previous research various machine learning method has been used to detect and filter the spam emails. The analysis method to detect the efficiency of detection the spam mails the hypothesis framework deployed to test the working efficiency of machine learning algorithm. The sentiment analysis which is based on detect the personality recognition which extensively used the email account, various hacker and spammers target that audience which does not know about the spam email and spam short messages. (W. Peng, 2018,) Enhancing the naïve bays spam filter through intelligent text modification detection, since the issue exist to detect and clean the spam email at the enterprise level, various industries application using the corporate email server to store and retrieve email data on particular server computing, so the email server does not able to detect the spammer. The email correspondence has been increase so far due to all official work required email communication which needs some tools and technique to do the various official task and need approval, so at this level urgent need of spam filter that detect, spam and eradicate it automatically.

Spam email stores in user inbox, and also change the text of original mails due to malware and phishing attack served in email system. Currently the naïve bays machine learning algorithm

effectively works very well to detect and refine the spam email from the user inbox. The implementation of novel algorithm that works with naïve bays which produces highest accuracy result that correctly detect the spam and ham. This project is going to use python programming framework which is most popular programming platform of machine learning and deep learning algorithm development discovery.

Python extensive works very well to detect various machine learning discovery, data science, big data analytics and artificial intelligence taking initiative to develop various AI based python libraries which essential works very well comparative to other programming language.

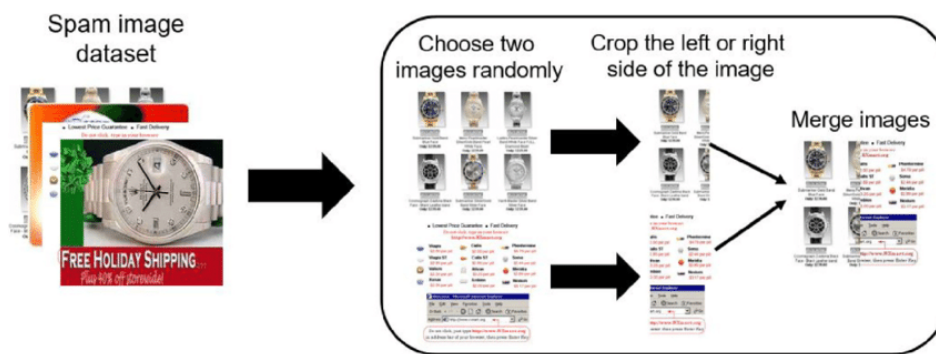


Figure 3: data augmentation process for spam image detection

Spam Email Detection Using Machine Learning & Neural Networks

(Sethi Sumesha, et al., 2021) Spam emails known as junk email which does not need approval to enter in the user inbox, the spam messages contains with spam hacker id that enter in the corporate to get the secret information and important asset details. The area of spam detection which is considered to be very limited due to some limited types of domains and networks.

Gmail is considered as one of the best email service in the world, it also helps to provide domain & network facility to enterprise business. Google built in with enhance cyber security features that support various domain and cloud computing also. Since the Google networks inspires from Yahoo, but the server and searching index of Google contains various types of security level it also support the Gmail to protect the end to end encryption for message sending and message

receiving to the user. Particularly machine learning method extensively used to detect the spam and ham emails, there are mainly two features has been introduce such as stopwords and wordcount these method analyze the email either the email is spam or ham.

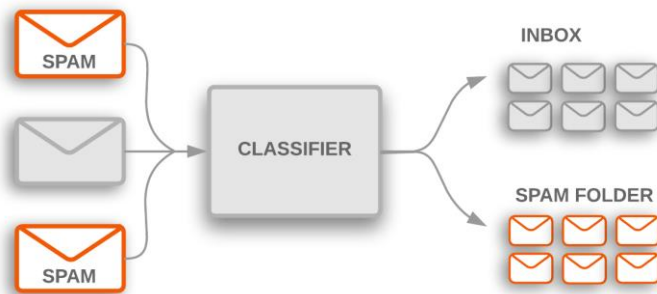


Figure 4: Spam Classifier

The entire email classification process is called spam detection which used to protect the user personal information and organization data. Various machine learning algorithm works efficiently to feature the spam and ham emails, various spam email contains malware attack link that used to destroy the organization server and networks computing and it also destroy the confidential data. Machine learning algorithm multinomial naïve bays, logistics regression, linear support vector machine including artificial neural network algorithm that used to determine the spam emails. Dataset has been used in this project from kaggle and Github repository. Detection of spam and ham emails and spam text and it detect the malware emails, malware emails used in previous windows releases, but the window 10 operating system updated.

Machine learning algorithm on email spam classification

(Neha., 2020) The message sending towards the user computer to other is considered sensitive matter due to some security breaches. Email classification method used with machine learning algorithm that works programming platform to manage the spam and ham messages over the internet usages. The research is focused on email spam detection beside this mobile communication short messages is also discussed and corporate large file text included with spam text has been highlighted. Unwelcome email coming from hackers site, it may use to spoil the communication network and destroy them. The development of deep learning approach with

neural network that effectively works to detect spam image by using the image dataset of spam image and clean image dataset. Various dataset has been provided on Kaggle and Github repository which support various research to deploy spam filter method to avoid any issues and problem in the organization. The objective of this research to conduct compressive machine learning algorithm technique such as SVM, DT, KNN, NB, LG, RF, LR CNN which combines all of these method in this research that support the research to deploy the email spam detector algorithm required in premises.

The rest of the thesis planned as follows in chapter 2, the comprehensive literature review and related work has been conducted, which is known as secondary research data, in chapter 3 the methodology framework of this research is planned and discussed including research paradigm and research philosophy, and the methodology of machine learning algorithm that perform the spam detector which provides in chapter 4 the analysis of machine learning algorithm such as SVM, DT, KNN, NB, LG, RF, LR CNN also evaluated with their efficiency and level and various spam filtering technique is also evaluated and critically analyzed with other researcher. Chapter 5 offers the discussion in which the frameworks of the overall structured machine learning algorithm benefits, advantages, drawback discussed, future requirement and challenges is highlighted. And at the end conclusion is described.

Aim & Objective of this Research

- The objective of this research to combine all machine learning algorithm to detect spam email and spam text and spam messages in the dataset.
- The aim of this research to develop python programming applications which executes on server platform and helps the organization to detect and filter the spam emails and spam text.
- The goal of this research used the primary data from Kaggle & Github repository, and the goal of this research to protect the confidential email of the user.
- The main goal of this research to block unwanted emails, block unwanted spam links and block malware phishing attack emails.

Chapter # 2: Literature Review

Introduction:

Literature review used to conduct the reviews of previous studies about the why, where and when it could be done, what are the major drawbacks in the previous studies and what are the advantages and disadvantages over the studies schemes. Literature review considering is part of secondary data collection method, it observer to highlight the research problem, research gaps, research area of interest to be happened on specific area and topics.

Related Work

Email spam filtering method used to review the filtering method but due to lot of spams in the junk folder, so the Gmail considered is specialized spam detector that separate the spam mails in the junk mail folder, and it automatically deleted after some period of time. (Gbenga & StephenBassi, 2019,) Machine learning for email spam filtering the method effectively works comparatively the email platforms, the issue is that the spam mails generates problem in the businesses and it difficult to tolerate in the daily usages of mails. The unwanted email considered as spam mails but it is very difficult to recognize that this mail is spam either or not in the junk of pool of mails, Microsoft outlook express mails does not handle spam filtering method due to poor framework of mails system. In recent years the development of machine learning method has been used to filter out the spam mails from the inbox as well as from the junk mail box. The application of machine learning method has been used by famous emails server such as Gmail, Yahoo, and outlook express, but the Gmail machine learning prediction method comparatively works well. The drawback of machine learning filtering exist in previous research but the replacement of deep learning model considered the best approach with the help of artificial neural networks layer. (Hanif Bhuiyan, 2018) Existing spam filtering method uses the machine learning method, email considering is most powerful and secure communication mechanism in the business dealing and serves the businesses over the world of commination. The increasing of data on daily basis detected through machine learning and automatically filter the unnecessary data from the emails server. various spam filtering method used in recent years

to detect & filter the spam messages such as knowledge based method, clustering method, learning based method, and heuristic process. Machine learning method used such as Naïve bays, SVM, k-nearest neighbor, bays additive regression method, KNN tree method, python programming recognized one of popular framework to handle this problem.

(Jain & Bhargava, 2021) SPAM filtering using the artificial intelligence technique such as artificial neural network approach included with deep learning method approach considered is best approach, it works on millions of hidden layers to identify the exact layer which one is considered is required approach needed to extract the exact data that need to be filter out. The antispam filter method used by Gmail, Yahoo, Outlook service these mails servers uses the existing machine learning models to predict the spam mails, but the issues exist sometimes it's not work efficiently, the arrival of hidden neural network layers change the strategies of machine learning prediction and its works efficiently comparatively to filter out the spams mails. Hidden neural networks layers approach works efficiently due to some hidden model building approach has been used. The algorithm detect, refine, separate and delete spams mails automatically after some period of time. (Dalkılıç & Sipahi, 2017) Spam filtering with sender authentication network, spam filtering method from the sender authentication that predict the senders in the mail list, it capture more spam mails, due to low resource difficult spam method has been implemented in recent years, spam mails is identify and filter in separate mail folder which effective way to identify them in separate folder. False positive and false negative method on the sender authentication list is help out to determine the spamming method. Sender policy framework protocol recognized well known origin method of spam filtering technique, due to sender authentication network developed to extend the spam filter recognize with sender IP address and network protocol , if the spam filtering method does not exist with comparison of false positive and false negative method, the way they are using to detect the spam on positive and negative layers mechanism of deep learning approach, result has been generated on false positive and false negative approach. (Aakash Atul Alurkar, 2019) Comparative analysis conducted to analyze the email spam classification technique using the machine learning approach. Gmail, Yahoo, Hotmail widely used by billions of people on daily, which connect all over the world people within a second. Machine learning approach has been widely used in all over the world, the detection of spam mails is detected on Google source platform in great efficiency the performance of mail sender and receiver has been recognized very well due to

powerful mechanism of Google API with deep learning approach. Machine learning supervised learning and unsupervised learning method recognized on train and testing the models, the retrieving method approach through the API layers of machine learning models.

(Wang, 2012) The optimal design method based on hierarchal spam filtering method based on greylisting, the threats on emails is going beyond with normal email service than ever before. Present spamming method is used to block the spam mails but the deletion method does not exist in this year, the recognition of spam mails that recognize large number so spams mails. The overall hierarchical method used with greylisting approach which based on CBFG which uses three layers of filtering method. (Yang, et al., 2017) Describing and predicting enterprise email reply behavior based on small data sample collected from surveys, the automated mails generated by the consumer on enterprise level of application businesses, the commercial mails that used to predict with evidence on personal account mails activity in the organization. The evidence of predicating the emails automatically reply recognized on various factors due to various factors that generate reply to the customer on consumer behavior activities but the modelling of automatically replies mails works efficiently on various matter to generate various replies to the customer on dealing of business products and financial recovery. Developing model to predict the automatically sender replies and how it will take long to take evaluation on it. The avocado emails collection describe the importance of automatically replies that based on thousands of avocado customer dealing on daily basis. (Hu, et al., 2021) The end to end encryption measurements on email spoofing attacks, phishing attack on emails recognized one of worst attack that hack email data due to sending spam mails and spam links which is dummy link that redirect on unknown website and steal the important user data. The cause of spam message attack on various organization enterprise email platform that does not contains particular security mechanism, email spoofing method uses to recognize the spam mails from the inbox. Studies show that the mails sender how is possible to scan the spoofing emails from the sender list. Penetrate mail box reach the defense on forged mails. 35 email providers tested the user reaction which was based on spoofing through the real world data on phishing email attack, email protocol uses to detect the spoofing such as Yahoo Mail, iCloud, Gmail uses the machine learning spoofing method to detect the unwanted mails from the inbox. Mostly the Gmail provides the end to end encryption that does not easily detect and hacked due to Google secure API layers comparatively Yahoo, Outlook does not provide the end to end encryption services.

(Bandaya & A.Mi, 2011) Analyzing the internet emails data through the method of email data spoofing, the spammers and phisher attacker readily availed to hack the enterprise emails data and get profit to send these data to other competitors. The above spam email to keep on top to generate the recipient mail box which aim to maximize the chance of action to detect & recognize the unwanted emails. The spoofing method can cause the confusion to recipient and it creates problematic situations over the client and customer to dealing with the kinds of data. (Sahni, 2021) Analysis of naïve bays algorithm uses in email spam filtering reviewing with the existing method approach that analyze the strength of machine learning prediction method as well as deep learning models prediction that ideally recognized as best method to detect and separate the spam messages and delete then particular time period. Author recommended that the approach of deep learning method is ideally works on predict the spam filtering method. The deep adversarial learning method effectively works on spam. (M., et al., 2016) email spam detection possible with using the best programming approach such as python machine learning models that predict and analyze the spam mails from the Gmail, Yahoo, Outlook express server. In these days all official correspondence possible through the emails service that communicating the official's emails. The problem exists on internet due to spam email messages that used to waist the employee's times due to unintentional detail and attraction of images and offers. Spam mails that uses unnecessary space of email server, and cause problem customer mails. The spam mails is mechanism of email phishing attack which used by the various industries to steal the important information of the organization and company asset details and other financial secret report. The machine learning approach with TFIDF algorithm term frequency inverse document approach implemented with support vector machine algorithm, result has been compared with confusion matrix of support vector algorithm with obtain the resultant accuracy of 99.9% of trained data and 98% testing data that obtain through the term frequency inverse document and support vector machine algorithm.

(CHARAN, 2021) Email classification using with machine learning prediction method ideally recognize the best approach due to increase number of unnecessary email on daily basis that increase the unwanted spams data on email servers and it uses to slow down the working of emails server. So there is need to creates an antispam filter that does not store any spam email in the inside the email inbox. In recent studies the development of machine learning spam filter widely used to detect and filter the spam emails. Machine learning based method effectively

works to recognize the spam emails spam users it effectively works on Gmail, Yahoo, Outlook express to identify the any spam activities and spam email in these brands. The results shows that the existing machine learning method does not works effectively that requirement. So the achievement of deep learning method and deep adversarial learning method dealing with spam filter effectively. The future works required to works more efficiently to detect and delete automatically and does not stores inside the inbox and uses the email server memory.

(Gangavarapu, et al., 2020) 99% accuracy of machine learning spam detection of unsolicited bulk mails, machine learning model has been extensively deploy by internet service providers and email service provider such as Yahoo, Gmail, Outlook uses machine learning python codes to detect and delete the spam messages in the email server. There are several phishing email attack occurred due to cyber security issues in the internet service providers domain. Email phishing attack uses to send spam email to users and it sends dummy link to redirect the user to another website which causes to steal the important information from the email. Email phishing attack identified with ethical hacking, spam emails not only wasting the user times it also consumes lot of internet bandwidth. Spam mail also contain some malware attack information which operate in windows operating system, various window operating system still not secure due to firewall breaches and built in antivirus was not providing the secure services to hide the windows data and files. The spam mail messages contain spam attachment which execute on window directory system 32 and slow down the working of operating system, so window technology still not well due to less security features in it. (Dada, 2018) Additionally the logistic model tree introduction machine learning method for email spam filtering, email spam uses to changes the behavior of employees email and steal the essential information which uses to beat the competitors. Machine learning approaches uses to solve the spam problem in most recognize internet service provider such as Yahoo, Google, Outlook server uses various machine learning and deep learning based technique to solves this issue in organize way. The conventional method of machine learning uses the black list strategies to block the spam ID over the internet services. IP addresses, domain services, was not able to identify the phishing threads so the programming interface uses the spam filtering and removing method to organize the enterprise application.

(Dada & Joseph, 2018) Random forest machine learning method for email spam filtering which is recognize best prediction method to spam emails. The traditional method widely uses in spam filtering such as black and white list uses the domain IP addresses which effectively works to

detect the spam list and put it on black list. The high performance and reliable spam method which approach to successfully detect the email spam messages which detect the spam emails. The proposed random forest machine learning method which effectively works to detect and filter the spam message, objective was to detect the spam email from emails, the prediction accuracy with less number of features from the Enron public dataset contains 5180 emails on both ham, spam and normal emails. The set of prominent emails featured with extracted the apply the random forest algorithm method of machine learning with 99 % accuracy with false positive and false negative rate uses the weka data mining prediction software. (Shirani-Mehr & D. Delvia Arifin, 2016) Enhancing method to detect the spam email messages and mobile message the detection of mobile phone short messages service perform using FP-growth and Naïve bays classifier, the short messages services is still very popular due to communication medium sending services. SMS filtering performance by combining the two data mining task associated with classification and FP-growth. Utilizing the data mining approach in frequent patterns of SMS and naïve bays classifier algorithm is whether to identify the message is span or ham, using training and testing the sms dataset the naïve bays and FP growth produces the highest detect accuracy of 98.5% with improved precision score.

(Anon., , 2017,) SMS spam detection using the H2o framework, various people receives spam messages over the mobile telephone which causes the problematic situations. SMS spam method already exist in previous research but the improved method requires to improve the performance of the detection method. Support vector machine and naïve bays algorithm and many other machine learning algorithm works to refine the working of email spam filtering which used to recognize the method of accurate F1 score and with highest rank of confusion matrix detection over the spam detection. The comparison between random forest and deep learning method uses to identify the various detection approach and accuracy of spam detection has been addressed in this studies. Result shows that the URL based text messages contain highest probability of detection due to internet API layers. The dataset uses from UCI machine learning repository, experiment results show the faster detection method spam detection, comparison of deep learning method contains highest scores random forest generate highest score with 50 trees and 20 depths including precision recall and F1-score 98% 96% 99% efficiency of detection. (NurAmir, et al., 2019,) SMS spam messages detection using the term frequency inverse document and random forest tree algorithm deploy on sms spam messages data collection which

is collected from machine learning UCI repository and kaggle platform. But the experiment results show that the detection of spam filtering accuracy is about to 97.50% accuracy of random forest tree algorithm.

(Sagar & Rutvij, 2020) In dealing with security and privacy concerns the machine learning uses the best prediction method to identify predict classify the security mechanism. Machine learning uses real time attack detection method with various API of python library packages to identify the real time problem and produces high quality accuracy. The real time decision making using the big data analytics approach is used to identify the approach of reduced cycle time of learning and cost efficiency including error free processing. (Onashoga, et al., 2015) An adaptive and collaborative server side sms spam filtering method with artificial immune system, existing system in 2015 does not able to stop the spam messages due to the problem exists on worldwide due to limited bag of training and validation method. Artificial immune system based on innate mechanism, user feedback and quarantine and tokenize the schemes of immune engine. Novel English corpus consisting of 5240 sms offering various studies including results show that the sms data with spam keywords. The 97% accuracy of artificial Bayesian with client side scripting to detect and spam sms and filter out in the mobile phone reserve user. (Islam & Ahmed, 2019) Social networking sites provides the Bangla text content through YouTube, the English text and Bangla text content combine schemes uses to spread the spam sms over the mobile phones users which is very difficult to identify them, the machine learning prediction model detect with 84% of accuracy including 70% F1 score and precision call.

(Maqsood, 2018) Microsoft .net library was used to detect the spam sms detection using the machine learning approach, the short sms on mobile phones users is also creating meta data that also engaged the user with annoying activity, the spam messages is widely consumes mobile users to engage mobile device and halt the operating system due to insufficiency memory of processing and rom. Random forest algorithm developed with C# programming it classify the dataset in terms of spam and ham. (Douzi, et al., 2020) Hybrid email spam detection model using the artificial intelligence neural network deep learning method to detect, separate and remove from the email. Its works very efficiently comparative other machine learning algorithm. The increasing volume of spam emails generated with essential of exact anti-spam filter detection method to detect unwelcome mails. The unwanted emails coming from other sources to beat the

competitors and extract the secret data of the industries and also uses to steal the window operating system files and change with malware files. Author propose the hybrid model uses with artificial neural networks paragraph and vector distributed memory. The vector distributed model built the spam mails in background of resource to identify the spam sender. Conducted empirical research to identify the spam email problems including machine learning method.

(sewayaa, et al., 2021,) the most accurate spam detection proceed with the machine leaning method which consider the accurate spam detection method, the algorithm of natural language processing which uses to analyze the email text and sender id that effective way to identify the spam text and spam sender. The machine learning natural language processing algorithm works on text mining to mine the text with lower case, upper case, special character including comma separated values. NLP algorithm able to identify the text within the spam and real messages. It also indicate the sender is fake or real. Author combines other method such as SVM, naïve bays, k-nearest neighbor, logistics regression, decision tree, random forest tree with accuracy result of 99% of spam mail detection (N. Kumar, 2020,) email spam basically in these days common problem for the every organization to meet this challenge various machine learning algorithm including deep learning logic to avoid spam mails. Email phishing attack sent via using the spam mails, the spam contains with dummy link and it redirect the user to another link and steal the information. For creating fake profiles, fake emails on social networking sites, it easily capture by the hackers and used for spam purposes. Phishing attack used in enterprise organization, spam user does not identify, mostly the spammers target the user which does not know the spam mails attack, the spammers objective is to commit fraud and steal essential information of the user, it might be used for human trafficking, so this thesis going to implement the spam email and spam user detector using the machine learning algorithm, so machine learning algorithm deploy in this work effectively to detect and filter the spam mail dataset with effective accuracy rate and precision call.

(Jun, et al., 2020) Spam detection approach used to detect the secure mobile message communication using the machine learning algorithms, mobile message communication is still insecure due to spam messages and spam calls. So the problem has been solved to deploy the machine learning spam detection method. Machine learning algorithm has been widely used by

various researcher to meet spam messages, spam calls, spam email challenges. Machine learning classification algorithm implements to detect, clean and filter spam emails and spam messages such as K-Nearest neighbor, decision tree, logistic regression are used to classify the ham and spam messages in mobile device communication. The SMS spam message dataset is used for testing the method, further the dataset is divided into two parts such as training and testing purposes. The outcomes shows with absolute 99% accuracy for using KNN, DT, and LR. (Santoso, 2019) Additional the analysis part has been also tested with to verify the algorithm performance to detect and identify the machine learning email spam detection performance in the analysis phase accurately. Spam emails is very annoying email that take attention towards the other side and wasting the time of the person. Very few companies email accounts not the email server provides the feature of spam detection, email does not contain the facility to identify the spam id and spam messages. Network administrator separately added the module in the server to detect and filter the spam emails so the implementation workout to secure the user from the spam emails. Logistic regression, decision tree, random forest tree successfully deploy in separate module of the email server to detect the spammers and the efficiency requirement has been tested to deploy the machine learning algorithm in the email server. The training and testing part of spammers is also tested to deploy the machine learning method which proceed the highest detection accuracy score rate about 98%.

(Govil, et al., 2020) In today world, everything is considered as data and world of data science including artificial intelligence method. Rapid growing of data with rapid amount of email correspondence has been increased on daily basis. Widely used emails are commercial so these commercial emails might be able to spread email phishing attack with malware. So this time the malware execute in the system and slow down the memory and execute continual tasks and system does not respond towards the user requests. So this research show that the development of machine learning spam detection algorithm which used to separate the real emails and spam emails.

(Hirano & Kobayashi, 2019) Machine learning based ransomware detection using storage access patterns which is obtained from live-forensic hypervisor. The rapid increasing of internet of things devices including cloud service with cyber physical system, various cyber security attacks on enterprise and public sector, beside this ransomware attack damaged the UK National Health

Service and various enterprises in the year 2017. Various researchers suggest solution to handle ransomware attack with prevention system. So to solve this ransomware attack problem the machine learning method has been used to avoid the ransomware attack problem. Ransomware attack samples data has been collected from forensic science center and it extensively used for testing purposes, the researcher first obtain the storage access patterns by using the samples of ransomware attack data hypervisor called waybackvisor. The suggested solution of ransomware attack to develop the machine learning algorithm such as K-Nearest, Support vector, random forest, decision tree with F-measure score about to 98% success rate. (Kumar, 2021) The novel machine learning method by using the emerging machine learning algorithm, machine learning is set of pattern that proceed to computer capable to learn without any programming effort. Machine learning support to validate the approach to support information. Machine learning algorithm works very fine to detect and clean the malware attack from the system. The functionality of machine learning method is used to detect the file either the malware is contain or not.

(Kumar, et al., 2018) Deep learning based image space detection various hackers and spammers employing to detect the obvious activity with spam image detection, spammers might be change the image and might be the original image fool the user, in this research the convolutional neural network method has been used to detect the spam image including the dataset of image 810 natural image and 928 spam image the classification accuracy achieved with about to 91.7%. (Akinyelu, 2021) Various advance method widely used to detect spam email messages and web spam detection, including social network spam various approaches has been widely used to inspired the working of email spam detection, mobile sms spam detection, web spam detection effectively detect filter out the spammers and hackers. (Singh, et al., 2021) The approach of email spam detection using the deep learning approach works efficiently by using the deep learning algorithm such as convolutional neural networks detect the image of spam messages by using the image dataset, beside this machine learning method used to detect and manage the spam email messages by using the Naïve bays, SVM, Decision tree, K-nearest neighbor algorithm and logistic regression method extensively used to detect the spam emails including the accuracy about 95% to 99% with successful F1 score and confusion matrix.

(Sharif, et al., 2020) Deep capture method used with deep learning method effectively to detect the spam email images and real images by using the image dataset. The dataset used from Kaggle repository and Github repository. Data augmentation possible to deploy the method of deep learning approach that effectively used to detect the real and fake emails. The convolutional neural network method used for detection of image data in the python library. The deep capture using the CNN algorithm which is part of deep learning algorithm that capture the image into billions of hidden layer with neural network. Neural network layer is construct with billions of hidden neural layer. Input layer, hidden layer, processing layer, output layer. These hidden layers match the pattern of image with evaluate to justify the exact pattern of image and get the exact result that required by the researcher. CNN network works on pattern of billions and millions of hidden neuron, the neural network architecture capture from the human brain like architecture that capture and make process the exact image by using the real world dataset and processed in more organize manner.

(Wu, et al., 2017) Twitter spam detection based on deep learning, proposed a machine learning based method to detect the spammers using the blacklisting method to detect the spamming activities on the twitter site. Recent method achieve 80% accuracy so the machine learning based method does not produce accurate result in 2017, so blacklisting method does not able to capture the spammers that required for accuracy. The novel method based on deep learning algorithm that the syntax of each tweet learned through word vector training mode, then proceed to binary classifier which was based on presentation of dataset. The text based detection method detect the fake and spam tweets with activity of the twitter account, this method analyze the twitter data which based on text. (G. Chetty, 2019) Deep learning based spam detection system developed the model which is based on combination of the word embedding technique including neural network algorithm, the word embedding method allows that distribute the presentation of words. Deep learning method used to learn the feature of text within the document and present the embedding space and these method used to learn the text, the deep neural network analyze the text from the text document and separate the spam text in large corporate document at the enterprise level.

Conclusion of Literature Review

In this studies the comprehensive literature has been used to analyze and predict the method of spam email messages detection including with spam email detection. Beside this spam mobile phone message data is also present in this studies, spam text document which is extensively used at corporate business level. Spam email that was issues of 2 Era, still the issue completely not resolved but the spammers and hackers are blacklisted in the email. The aim of this secondary data is used in this thesis and also used primary data in this thesis to detect and filter the spam email using the python machine learning method. This thesis combines all of the machine learning method and present all of the spam email detection method in one research. Previously the works is not present accurately due to some incomplete library of data science and machine learning. This research method is going to detect and filter the spammer and spam emails in more effective way.

Method Used:

1. Machine learning method used to predict filter and remove spam email by using the spam email dataset.
2. Machine learning Algorithm such as Naïve Bays, Support Vector Machine, Logistics Regression, Random Forest Tree, Decision Tree ,K-Nearest Neighbor.
3. Deep learning method to detect spam images from image dataset Convolutional Neural Networks.
4. Running Python application that detect spam email and present the outcomes of real emails.

Chapter # 3 Methodology

Introduction

Research methodology invent to build the machine learning and deep learning algorithm, so the algorithm helps a lot the email spam detector method which will help in the entire domain of organization and enterprise level in which email server, domain server has been managed.

Research studies carries to conduct secondary research data and analyzed various machine learning and deep learning approaches developed. Primary data is collected from Kaggle and Github repository.

Research Philosophy

First email spam originate in 1978, created by Gary Tuerk, worked as marketing manager in digital equipment company. The first spam message included with to sell a computer, that was time only 300 internet users exists that part of ARPANET. After this discovery the spam starting to send people happy birthday messages in the email inbox. (message, 2021). Email spam now considered as junk email refers to unwanted and unwelcome emails that automatically stores in junk folder inside the email inbox. Email spam messages sent to the large audiences with large number of emails list (spam?, 2021) , usually it was sent from botnet which is network of computer that infected with malware already and it was controlled by single community the attacking party recognize as bot herder, spam would be sent from social media sites and distribute them to randomly generated email id that target the exact enterprise audience.

Various people received spam messages but unable to reply them because it contains spam images, spam links, spam text and sometime might be contains malware attachment, the malware attachment download in windows and stop the internal working of operating system. Email spam senders only financial motives, spammers sent fake attractions deceive recipient by offering dummy products discounts and celebration tickets etc. famous spam subjects & message includes, pharmaceuticals, adult content, financial services, online degree, work from home jobs, online gambling crypto currency.

Data Collection and interpretation

Mostly the spam mails and spam message spam text dataset has been collected from Kaggle repository, it was also provided by the datawold site. The data collected from social sites and uploaded on Kaggle repository and it is primary source of data. (Dataset, 2021), the dataset folder contains spam folder and ham folder, each folder contains with emails. The email data is convert into a dataframe and written in csv file that easily readable and accessible to python programming server.

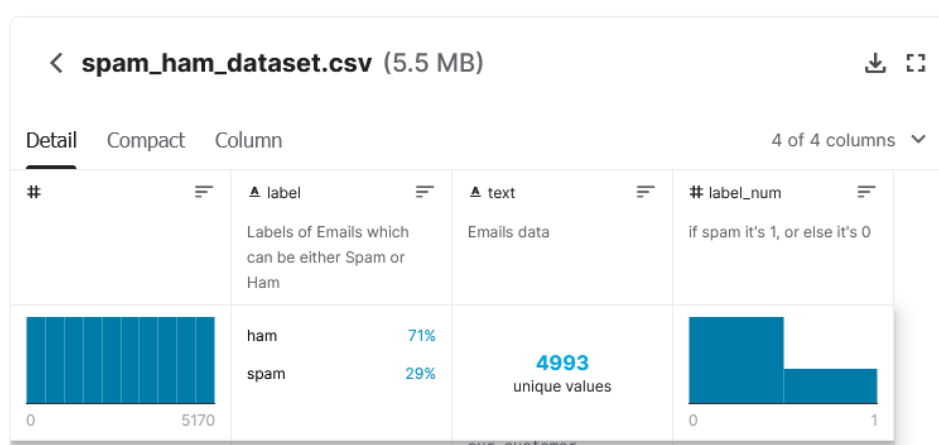


Figure 5: Dataset information

Research Strategy

Gmail Spam Filter:

Google data center develop various rules and policies to determine either the mail is spam or real, with this rule various statistical values connected with email server, depending on the outcomes of the statistical rules it present that it is spam. The weighted features built an equation and the test has been conducted on the basis of test result, threshold sensitivity decided on user spam filter to address that mail is spam or ham. (Gbenga, et al., 2019) Google data center uses state of the art quality machine learning algorithm to process Gmail which is known as spam detection machine learning algorithm which is based on logistics regression, and neural networks that categorize the emails. Neural network is branch and subset of deep learning algorithm which

exist in machine learning pattern. Gmail also deploy OCR service to deploy the emails that uses the Gmail user emails collected from image spam, machine learning algorithm also works on Google search engine to classify the Gmail email servers and predict and analyze each user email and take decision on values on textual based and image based either the email contains spam information and real information. (Gbenga, et al., 2019)Gmail rank with millions of classification method which improves the spam filtering strategies. Spam filtering works on principles of rules that depend on filtering settings that emerge the common algorithm discovery. Text based spam filter is difficult to recognize the real and spam emails.

Yahoo Mail Spam Filters:

Yahoo known as first free email providers before Gmail, but later Gmail gets more attraction & popularity due to distribution of domain services, as compare to Google datacenter to Yahoo Datacenter, Google is much better than Yahoo due to daily basis improved algorithm strategies. Yahoo used spam filter method to detect the spam messages contain with spam user id & requests. URL filtering email content & spam complaint from user. Yahoo used combination method to filter out spam messages that provides the mechanism of valid user from being mistaken for spammers, (Gbenga, et al., 2019)SMTP errors that generate SMTP logs. Commonly method used by yahoo such as blacklisting method and whitelisting method that automatically block the spammers without knowing the user. Beside this whitelist is based on very strict rules and policies that block the sender without bothering the yahoo mail user, and many spam filter automate the whitelist, the anonymous sender list has been verify from yahoo database, if history describe the user is spammers with spam activities history, if the spammer detect the message sent to recipient inbox and it added to whitelist.

Outlook Email Spam Filter:

Outlook email is part of Microsoft account, previous known Hotmail, live, outlook is collection of Microsoft office suite application which combine and automate the office suite application into the outlook express, and it binds the email server to Microsoft domain server. Outlook allows the user to store data on cloud server computing, such as Microsoft Azure server. Microsoft one drive account that companion of Google drop box, one box and Google drive.

Outlook used encryption policy to send the messages to user email account which works on password authentication policies. Outlook have own filtering method.

Data Collection Method

Primary Data:

Primary data has been collected from famous data provide repository which is known as Kaggle and Github, that the only source of primary data collection which based on email spams data and real email data. Primary data folder divided into sub two folder spam email data and ham email data.

Secondary Data:

Secondary data has been collected in this research which is based on previous research publications, IEEE Journals IEEE Xplore, ACM library, Science direct and MDPI research including Google Scholars with peer reviews journals articles with good impact factor.

Data Analysis Method

Machine learning algorithm has been used in this project including deep learning algorithm that used to detect and filter out the spam and ham email messages.

Python Anaconda IDE with Jupyter Notebook used for data analysis beside this Python Pycharm also used.

Naïve bays classifier:

1. Naïve bays classifier developed to detect the spam emails by using the primary data of email which contains spam emails and real email, it classify with real on 1 that is true and spam classify to 0 which is false. Naïve bays used python pandas, numpy and nltk library. NTLK library works to analyze the word text in the dataset, categorize the word with lower and upper case letter, sklearn to extract the naïve bays classifier to detect the spam

text. Dataset divided into two categories such as spam and ham data that process in python with dataframe library 70% of training data is labeled as spam and rest of the other testing data is 30% labeled as real email data. Probability of spam divided by discard text that equal to probability of word discard occurring on the sentence that are labelled as spam multiply by training data into the probability of getting the data into spam, number of document which are spam that divided by total number of training data, the discard text known as likelihood which tells the theorem that this text is spam or not according to the mathematical theorem of naïve bays, probability of spam text divided by total number of text data, after training & testing the dataset that divides into x and y variable using sklearn library and extract the multinomial naïve bays model. (detection, 2021)

SVM support vector machine classifier

2. SVM support vector machine classifier used to build the email classifier that predict the spam and ham emails from the spam email dataset and real email dataset. SKLEARN library used to import linear support vector machine model in python jupyter notebook and it build the model to produce the outcomes of spam email and real email. Training the dataset by using the make dictionary function. Training & testing the dataset that divides the data into 70% and 30% respectively and make predictions with sklearn library and then prediction on dataset. Training the feature labelled which present the text information that is spam and ham. And training the SVM and Naïve bays classifier to build the model. (Mail, 2021)

Natural Language processing using python with Logistics regression spam email classifier

3. Natural Language processing using python with Logistics regression spam email classifier model using sklearn and countvectorizer library that count the word in the text file including stop words and count word, the method of training and testing the text data by dividing them and assign them the variable interpretation of the dataset which build the machine learning models and predict the spam and ham emails. (python, 2021)

Convolutional Neural Network deep learning algorithm

4. Convolutional Neural Network deep learning algorithm used that is classify the spam images such as birthday, offers, gifts and promotional images that part of spam emails. Training the spam images and testing the real images and predict by using the sklearn library and make assumptions with artificial neural network layer that produces the result and classify the various images into the model building shapes and predict the image is spam and real. Keras model and keras layer, for building the pandas and numpy library and testing the images and produces the outcomes after validating the images into the frameworks. (detection, 2021)

Email Spam Classification App that build by using the streamlit & python

5. Email Spam Classification App that build by using the streamlit & python library that used to quickly generate the app that executes in internet browser. This application used with multinomial naïve bays algorithm that build with sklearn python library. Dataset used from kaggle and Github repository and divides them into 2 category. After building the multinomial model. The application is designed in GUI layer novel layer which is called streamlit library that build quickly prediction app and predict that the email is spam or ham. (python, 2021)

Quantitative Method

- Statistical machine learning method has been used in this research
- GUI application executes with the help of statistical model in the browser

(Raad, et al., 2010) various advertising companies used the method of email that distribute product in the form of email, but the email server considered this advertisements as a spam emails because there is no resource background and preferred email id approach, while the spam classified in the email server that categorized the email into two frames such as wanted emails and unwanted emails. Might be anti-spam algorithm sometimes wrong estimation but the development of dataset and build according to the requirements of the demands and it actually works to refine the email classify framework. Email marketing companies developed a marketing software that does not work accurately due to less programming modules in it, so

far the machine learning algorithm is not in email marketing software Gmail Yahoo outlook considered these types of email as spam email. It analyzed with machine learning models that used with various algorithm.

(Clayton, 2007) Email traffic can be optimized with statistical method that used the numeric digits data on internet traffic bandwidth and used the various classify algorithm, since 1978 the email correspondence known as only source of information sharing and data sharing platform that uses internet protocols with SMTP and IMAP POP3 mechanism to send and received emails, so various hackers and spammers targets the corporate enterprise audiences and spread malware emails to destroy data and confidential information, it just because of sake of some minor finance. Later this activity in 1986 considered as crimes and cybercrimes activities, it might be sometimes punished and charges to pay some cash to execute the person.

Results of Research

The research outcomes present with various machine learning and deep learning algorithm that predict and analyze the spam and ham data that evaluated with training and testing the email datasets. And build machine learning models, python numpy and panda's library used beside this sklearn and multinomial and countvectorizer used in this project. Beside this GUI application streamlit which is efficient model building program that build GUI which executes in browser.

1. Naïve bays classifier to detect spam emails & real emails
2. Support vector machine classified detect spam and real emails
3. Natural language processing with logistics regression build spam and real emails model and classify spam mails
4. Streamlit with python used to build naïve bays multinomial model to predict the spam and real emails in browser.

Python Numpy and Pandas SKLEARN, Multinomial, SVM_CLASSIFIER, Naïve_Bays Classifier & streamlit used.

Sources

The following sources used in this research

1. Google Scholars
2. YouTube
3. IEEE XPLORE
4. IEEE Conference Paper
5. IEEE Journals
6. ACM Digital library
7. MDPI Research
8. Python Anaconda IDE with Jupyter Notebook version 2021
9. Pycharm IDE 2021 version.
10. Kaggle
11. Github

Ethical Concern

This research is original research works, it is not part of any researcher and scholars, and myself finished effort to create python machine learning application which used to produces the email spam detection by using the dataset. This research does not harm any person and any researcher, not part of any beneficiary and stakeholders. This research present stat of the art python programming features that extensive used machine learning algorithm such as naïve bays, SVM, LG, CNN and Streamlit. Other research works is cited with journal name and author name, which is considered secondary source of data, primary data has been downloaded from Kaggle repository and analyze them in Github framework. This research project is going to help researcher to develop efficient machine learning and deep learning algorithm.

Chapter # 4: Results & Analysis

Machine learning & deep learning based modeling has been tested and verified the email spam detection, various emails contains junks and spams text and useless information. Machine modelling build to analyze the spam email method, Google emails used machine learning model in Gmail services that predict and analyze the spams emails widely in the globe. Gmail service is one of the best email service in the world that detect and filter the spam and ham emails and forward the spams mails into the spam junk folder. Python machine learning based modelling has been built to test the email spam data which is downloaded from Github repository.

Email Spam Classification App that build by using the streamlit & python

(Freydenberg & Kevin, 2020) Project based learning to make software application using the novel python based approach, which is effectively used to test and predict the software applications within the environment. Software application tested the machine learning models and predict the spam and ham news in the internet browser. Streamlit is creative python application which used to present the application into the web browser and analyze the various working of the python application. The interactive python streamlit package instantly creates the web application which used to test any machine learning and deep learning project.

Streamlit python library installation:

Open the python Pycharm 2021 version and installed the following libraries in the Pycharm library.

1. Pip install streamlit
2. Add the streamlit package in Pycharm community 2021 version.

The following code generate in Pycharm community edition.

```

spamDetector.py x
1 import streamlit as st
2 import pickle
3 from sklearn.feature_extraction.text import CountVectorizer
4 import numpy as np
5 from win32com.client import Dispatch
6
7 def speak(text):
8     speak=Dispatch(("SAPI.SpVoice"))
9     speak.Speak(text)
10
11 model = pickle.load(open('spam.pkl','rb'))
12 cv=pickle.load(open('vectorizer.pkl','rb'))
13 def main():
14     st.title("Email Spam Classification Application")
15     st.write("Build with Streamlit & Python")
16     activites=["Classification","About"]
17     choices=st.sidebar.selectbox("Select Activities",activites)
18     if choices=="Classification":
19         st.subheader("Classification")
20         msg=st.text_input("Enter a text")
21         if st.button("Process"):
22             print(msg)
23             print(tvpe(msg))

```

Figure 6: Streamlit Code

And building the spam detection model in the python anaconda jupyter notebook to verify the spam emails and ham emails.

Dataset information & Python file

The dataset and python model has been used from Github repository. (Github, 2021). The dataset contains the spam and ham email text which contain the csv file.

```

In [2]: data=pd.read_csv("spam.csv", encoding="latin-1")
In [3]: data.head()
Out[3]:

```

	class	message	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

Figure 7: importing dataset in python

Data Cleaning & Pre-processing

Dataset has been clean to remove the unknown values in the dataset. It also used to verify the message and class values of the dataset. Data.drop syntax remove the unnamed columns in the dataset. And remove the NaN values in the dataset.

```
In [5]: data.drop(['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], axis=1, inplace=True)
```

```
In [6]: data.head()
```

Out[6]:

	class	message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

Figure 8: data cleaning process

Converting the text data into the numeric values to check and verify the spam emails and ham emails from the dataset. Data class to convert the class columns to 0 and 1 values, ham = 0 and spam =1.

```
In [7]: data['class']=data['class'].map({'ham':0, 'spam':1})
```

```
In [8]: data.head()
```

Out[8]:

	class	message
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

Figure 9: dataset conversion

Machine Learning Library to build spam models

Machine learning library import in python jupyter notebook, which used to analyze the dataset.

The following library has been used to predict the dataset values, sklearn library and countvectorizer library contains rich information to process the text message and class values that is based on the 0 and 1 value.

```
In [10]: from sklearn.feature_extraction.text import CountVectorizer

In [11]: from sklearn.model_selection import train_test_split

In [12]: X=data['message']
          y=data['class']

In [13]: X.shape
Out[13]: (5572,)
```

```
In [14]: y.shape
Out[14]: (5572,)
```

```
In [15]: data.isnull().sum()
Out[15]: class      0
          message   0
          dtype: int64
```

Figure 10: importing libraries and removing null values

Training & testing the dataset

To split the 20% data for testing and 80% data to training and make assumptions to verify the text data is contains spam value and ham values which is tested on the base of class values.

```
In [18]: x_train, x_test,y_train, y_test=train_test_split(X,y, test_size=0.2, random_state=42)

In [19]: x_train.shape
Out[19]: (4457, 8672)
```

```
In [20]: x_test.shape
Out[20]: (1115, 8672)
```

```
In [21]: from sklearn.naive_bayes import MultinomialNB

In [22]: model=MultinomialNB()

In [23]: model.fit(x_train, y_train)
Out[23]: MultinomialNB()

In [24]: model.score(x_test, y_test)
Out[24]: 0.97847533632287
```

Figure 11: training & testing dataset

Building the Multinomial Naïve Bays Classifier Model

The naïve bays classifier which contains multinomial features that used to predict the numeric class which identify the 0 and 1 values either it contains to proceed the spam and ham values.

```
In [22]: model=MultinomialNB()

In [23]: model.fit(x_train, y_train)

Out[23]: MultinomialNB()

In [24]: model.score(x_test, y_test)

Out[24]: 0.97847533632287

In [33]: msg="You Won 800$"
data = [msg]
vect = cv.transform(data).toarray()
my_prediction = model.predict(vect)

In [34]: vect

Out[34]: array([[0, 0, 0, ..., 0, 0, 0]], dtype=int64)

In [35]: import pickle
pickle.dump(model, open('spam.pkl','wb'))
modell = pickle.load(open('spam.pkl','rb'))
```

Figure 12: Testing the model

After building the naïve bays classifier model with the help of pickle spam files and other spam generator to verify the naïve bays modelling. (Nance & Baumgartner, 2021) Machine learning used to analyze the working of natural language processing, various labelled machine learning approaches is based on binary classification which used to predict the text in numerical form to analyze the values is spam either ham. Training and testing the machine learning models which divided the dataset and verify each text content based on NLP prediction.

Results of Streamlit python Spam Detector in Browser:

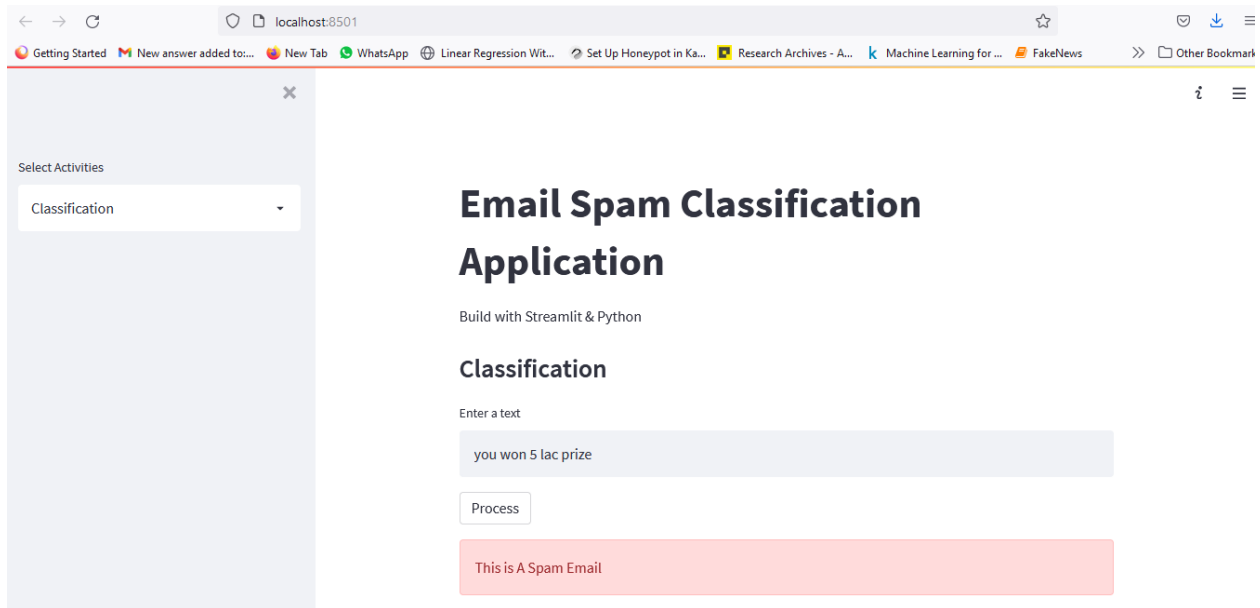


Figure 13: Email Spam Detector

The above mentioned application quickly build using the streamlit python library and make assumptions of spam and ham dataset. (Github, 2021)

Analysis:

(N. L. Octaviani, 2020,) The comparison of multinomial naïve bays classifier which used to predict the email spams, it also used to compare the outcomes of support vector machine as well as recurrent neural network, the abused email considered spam emails and it identify the sender identity, mostly it used for marketing purposes but it also used for making fraud, hacking data and steal information to send the malware information and dummy link to the user. The naïve bays model score is 97%.

Support Vector Machine Classifier

Support vector machine classifier (Shahi & Yadav, 2014) represent various feature to detect the spam and ham emails, the classification method used to predict the different domains such as text messages and email message which used from dataset.

Python Libraries:

1. Pandas
2. SKLEARN
3. SKLEARN Multinomial Naïve bays and Gaussian Naïve Bays
4. SKLEARN SVM

Data Set information

Dataset contains csv file downloaded from YouTube repository including Google Drives links

Python SVM model (Classification, 2021).

Split dataset in Training & Testing data

Split data in x and y variable dataframe. X = EmailText Y = Label the dataset.

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB, GaussianNB
from sklearn import svm
from sklearn.model_selection import GridSearchCV
```

##Step1: Load Dataset

```
dataframe = pd.read_csv("spam.csv")
print(dataframe.describe())
```

##Step2: Split in to Training and Test Data

```
x = dataframe["EmailText"]
y = dataframe["Label"]

x_train, y_train = x[0:4457], y[0:4457]
x_test, y_test = x[4457:], y[4457:]
```

Figure 14: dataset importing in python including testing & training process

Extract Features CountVectorizer:

(sklearn.feature_extraction.text.CountVectorizer, 2021) The method of countvectorizer is used to convert the text into the matrix of token counts. The dictionary of dataset is provided from the csv file dataset and it analyze the vocabulary size which based on analyzing the data, build analyzer () return the callable to process the input data. Build preprocessor return the function to preprocess the text before tokenization. Build tokenize return the functions that splits string into sequence of tokens.

Problem with Naïve Bays:

It assumes the attributes are independent and the weak point can be solved performing some statistical analysis before using the naïve bays to measure the correlation degree among features and then selection the most uncorrelated ones. Zero probability problem it can be solved by adding values one to the frequency each attribute used the Gaussian distributed method. It treats all attributes equally so some weights can be added to the important attributes to increase the contribution in the final decision, continues values attributes problem conversion from continues to discrete value is the solution of such problem.

```
##Step3: Extract Features
cv = CountVectorizer()
features = cv.fit_transform(x_train)

##Step4: Build a model
tuned_parameters = {'kernel': ['rbf', 'linear'], 'gamma': [1e-3, 1e-4],
                    'C': [1, 10, 100, 1000]}

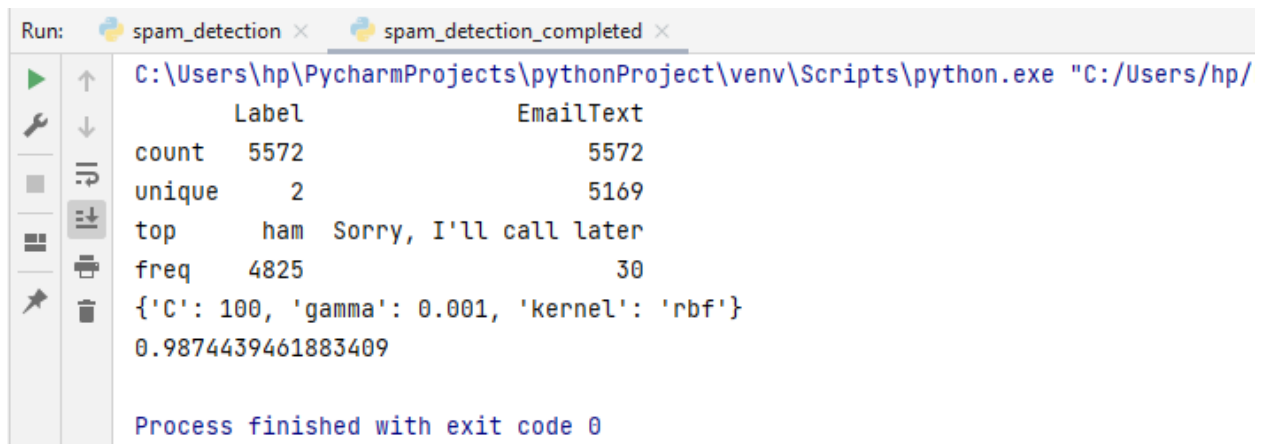
model = GridSearchCV(svm.SVC(), tuned_parameters)
model.fit(features, y_train)
print(model.best_params_)

##Step5: Test Accuracy
print(model.score(cv.transform(x_test), y_test))
```

Figure 15: Extract features & build the model

Results

The above result shows the SVM classifier accuracy about to 98% accuracy of spam and ham email detection which is achieved the best accuracy of the model. Label count the 5572 email text and unique values is 2 and top value is ham message which contains I will call you later. Frequency is 4825 and 98% accuracy of spam detection from the given dataset.



```
Run: spam_detection x spam_detection_completed x
C:\Users\hp\PycharmProjects\pythonProject\venv\Scripts\python.exe "C:/Users/hp/
Label      EmailText
count  5572      5572
unique    2      5169
top      ham  Sorry, I'll call later
freq  4825      30
{'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}
0.9874439461883409

Process finished with exit code 0
```

Figure 16: Outcomes of spam detection

Analysis:

(Roy, et al., 2017) The support vector machine classifier model used the artificial neural network to compare the performance of the machine learning models, the deep SVM model used to predict the ham and spam emails. The available statistical models which analyzed the text on countvectorizer strategies and convert the string text into the numeric digital which is analyze in 0 and 1 form. The SVM classifier executes with multinomial naïve bays model and built the model with effective python sklearn library and it build the 98% accuracy which considered good accuracy in the model building framework. Spam filtering method essential features in email detection which train and test the dataset and make the model predictions which is used the labels text dataset.

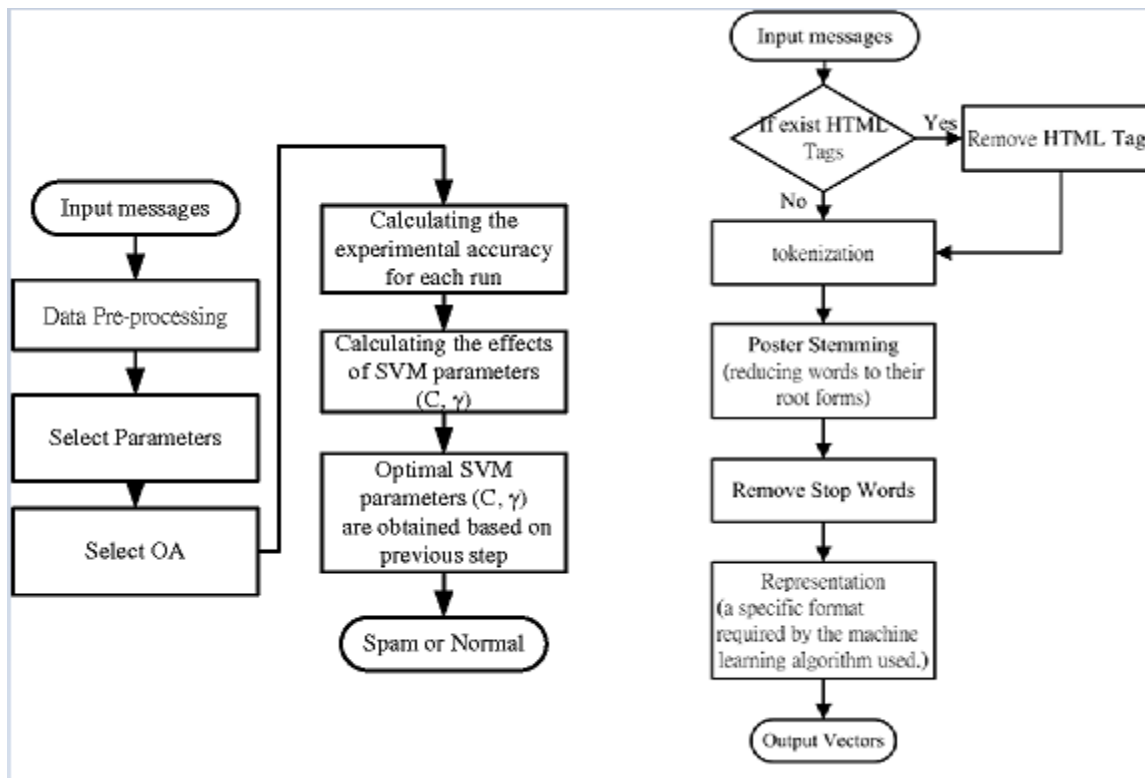


Figure 17: SVM Spam Detector flow chart

Naïve Bays Classifier to Detect Spam Emails

Naïve bays classifier is much effective machine learning algorithm, it uses multinomial method to extract the text countvectorizer and make assumptions on it, dataset split into two phases one is test and other training set. Machine learning python libraries used to import the spam and ham text algorithm which classify on binary method.

Importing Python Libraries

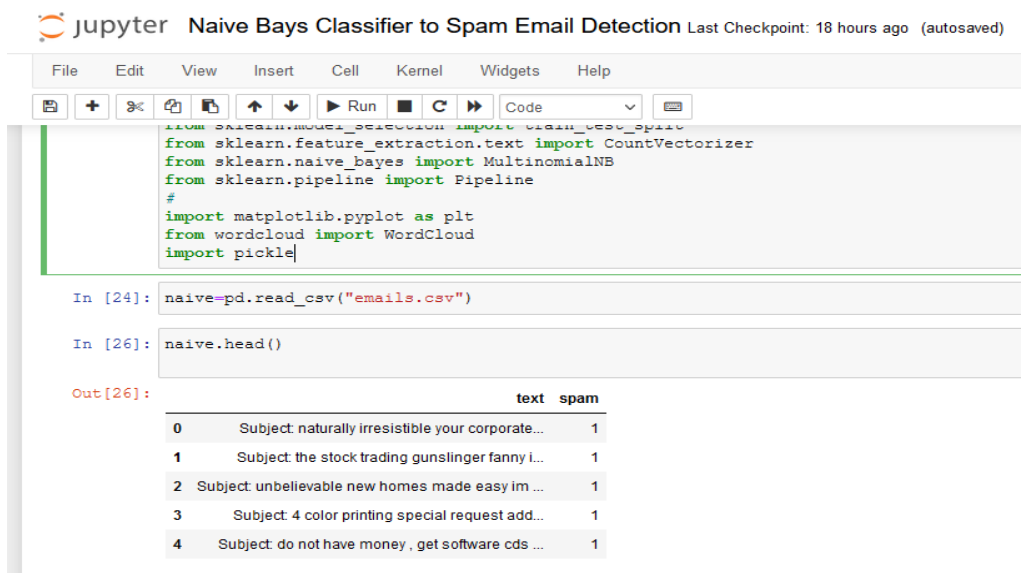
The following libraries import in anaconda jupyter notebook:

1. Sklearn Naïve bays
2. Sklearn multinomial
3. Import countvectorizer
4. Import wordcloud
5. Import pickle
6. Import pandas and numpy

Read the csv file email csv file

Data set information

Data set downloaded from Github repository, email csv file



The image shows a Jupyter Notebook titled "Naive Bays Classifier to Spam Email Detection". The notebook has a menu bar with File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. Below the menu bar is a toolbar with icons for file operations, running, and other functions. The code editor shows the following code:

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import Pipeline
#
import matplotlib.pyplot as plt
from wordcloud import WordCloud
import pickle
```

The output of the code is shown in the output area. It displays the first five rows of the dataset, which is a CSV file with two columns: "text" and "spam".

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny l...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1

Figure 18: importing naïve bays library

Date Preprocessing

Data preprocessing has been performed to analyze the dataset frequency, count the word count and count the description of the dataset levels.

```
In [27]: naive.groupby('spam').describe()
```

Out[27]:

	count	unique	text	top	freq
spam					
0	4360	4327	Subject: * special notification * aurora versi...	2	
1	1368	1368	Subject: naturally irresistible your corporate...	1	

Figure 19: group the dataset

Training & Testing the dataset

1. Split the dataset, split into x and y variables

```
In [30]: X_train, X_test, y_train, y_test = train_test_split(naive.text, naive.spam, test_size=0.25)
```

```
In [31]:
```

```
#Count vectorizer example to convert text to numbers.
corpus = [
    'This is the first document.', #1
    'This document is the second document.', #2
    'And this is the third one.', #3
    'Is this the first document?', #4
]
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
#unique words in each of the four document above. We have 9 unique words.
#We can take these words and treat each as a feature, as a column.
print(vectorizer.get_feature_names())

['and', 'document', 'first', 'is', 'one', 'second', 'the', 'third', 'this']
```

Figure 20: countvectorizer conversion of text

Building the text matrix with the each word by using the countvectorizer to represent the words as counts to see the individual features of the text.

```
In [32]:
```

```
#Build a matrix with each of the word. This represents words as count.
# We can use these individual features
print(X.toarray())

[[0 1 1 1 0 0 1 0 1]
 [0 2 0 1 0 1 1 0 1]
 [1 0 0 1 1 0 1 1 1]
 [0 1 1 1 0 0 1 0 1]]
```

```
In [33]:
```

```
#Count Vectorizer to convert words into a matrix of features.
v = CountVectorizer()
X_train_count = v.fit_transform(X_train.values)
X_train_count.toarray()[:2]
```

```
Out[33]: array([[0, 2, 0, ..., 0, 0, 0],
                [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

Figure 21: building the text matrix

CountVectorizer to convert words into matrix features

```
#Count Vectorizer to convert words into a matrix of features.
v = CountVectorizer()
X_train_count = v.fit_transform(X_train.values)
X_train_count.toarray()[:2]

Out[33]: array([[0, 2, 0, ..., 0, 0, 0],
                [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

Figure 22: countvectorizer transformation

Multinomial Naïve Bays algorithm to classify text and build the model

Multinomial naïve bays algorithm used to predict and classify the text either the text is spam and ham. It convert it into the string text into the numeric digit and perform calculations. 1 for spam and 0 for ham

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Posterior ← $P(A|B)$ Likelihood ← $P(B|A)$ Prior ← $P(A)$ Normalizing constant ← $P(B)$

$$P(B) = \sum_Y P(B|A)P(A)$$

Figure 23: Theorem

Results

```
In [34]: #We are using Multinomial Naive Bayes model to classify text. For this example we have count of each word to predict
# the label ham or spam.
model = MultinomialNB()
model.fit(X_train_count,y_train)

Out[34]: MultinomialNB()

In [35]: emails = [
    'trading limit and policy changes vince - here ' ' s a summary of what ' ' s going to the bod , along with updated policy
]
emails_count = v.transform(emails)
test = model.predict(emails_count)
for i in test:
    if i == 0:
        print("ham")
    elif i == 1:
        print("spam")
ham
```

Figure 24: Multinomial naïve bays classifier

Naïve bays model accuracy score represent the accuracy level about 98% of spam email detections. Get the model accuracy score and save the countvectorizer file.

```
In [36]: #Get the model accuracy score
X_test_count = v.transform(X_test)
model.score(X_test_count, y_test)

Out[36]: 0.9888268156424581

In [37]: #save the count vector file
pickle.dump(v, open('count_vect', 'wb'))

In [38]: # Saving the model
pickle.dump(model, open('email_class.pkl', 'wb'))
```

Figure 25: model accuracy score

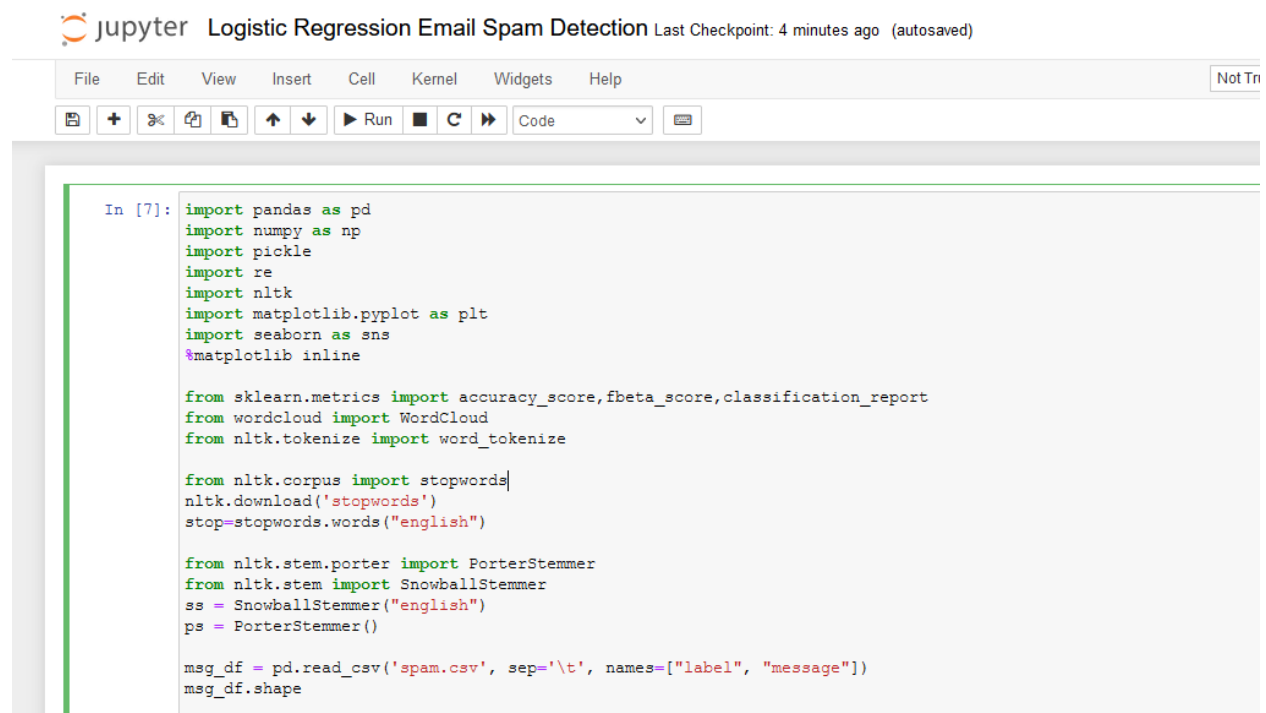
Analysis

(Kumar, 2018,) Integrated naïve bays spam detection approach has been used in this project the classify the spam email detection the particle swarm email detection approach has been carried out to optimize the good accuracy, dataset clean and stabilize with preprocessing approach, after training and testing the email dataset, multinomial classifier has been used filter the ham and spam emails.

Machine Learning Logistics Regression to Detect Spam emails

Best model to handle text classification problems, Based on binary output representation
Involves probabilities of both outcomes. Process spam training data, and process non spam training data, process spam testing data and build the probabilities model to predict spam and ham emails.

Data Preprocessing and Importing CSV



The image shows a Jupyter Notebook window titled "Logistic Regression Email Spam Detection". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for saving, running, and other actions. A code cell is active, containing the following Python code:

```
In [7]: import pandas as pd
import numpy as np
import pickle
import re
import nltk
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

from sklearn.metrics import accuracy_score, fbeta_score, classification_report
from wordcloud import WordCloud
from nltk.tokenize import word_tokenize

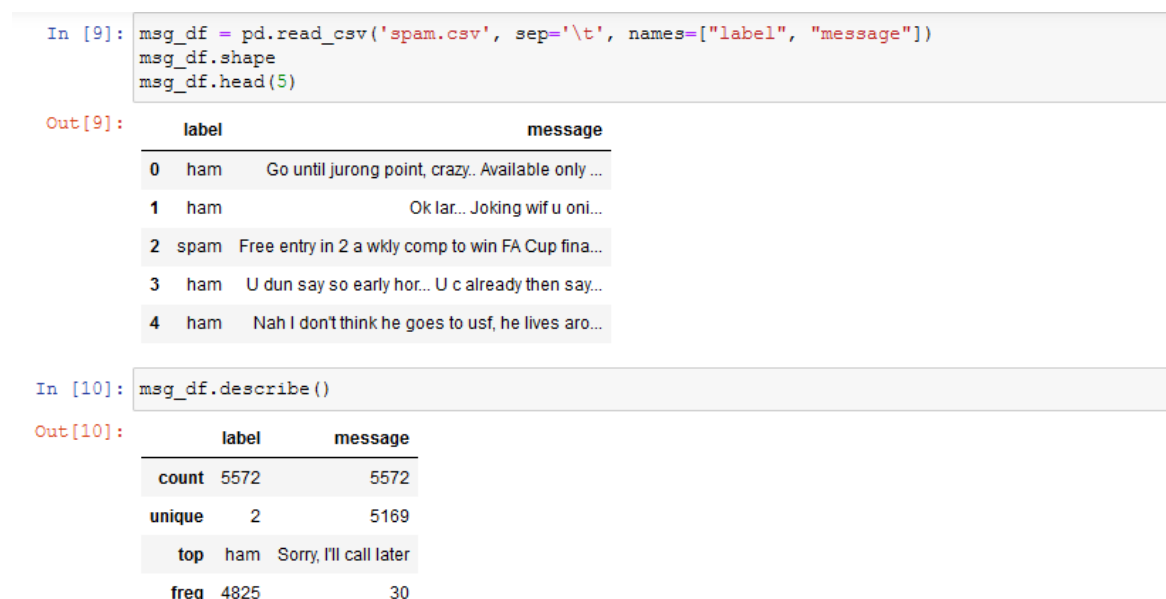
from nltk.corpus import stopwords
nltk.download('stopwords')
stop=stopwords.words("english")

from nltk.stem.porter import PorterStemmer
from nltk.stem import SnowballStemmer
ss = SnowballStemmer("english")
ps = PorterStemmer()

msg_df = pd.read_csv('spam.csv', sep='\t', names=["label", "message"])
msg_df.shape
```

Figure 26: logistics regression

Verify the spam email dataset by visualize the columns values in the dataset. Detect the statistical values of spam email messages.



The image shows a Jupyter Notebook interface with two code cells and their outputs.

Code Cell 1:

```
In [9]: msg_df = pd.read_csv('spam.csv', sep='\t', names=["label", "message"])
msg_df.shape
msg_df.head(5)
```

Output 1:

	label	message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

Code Cell 2:

```
In [10]: msg_df.describe()
```

Output 2:

	label	message
count	5572	5572
unique	2	5169
top	ham	Sorry, I'll call later
freq	4825	30

Figure 27: map the algorithm and count the text values

Text values count of spam and ham emails present the actual text wordcount within the dataset.

```
In [12]: msg_df["label"].value_counts()
```

```
Out[12]: ham      4825
spam       747
Name: label, dtype: int64
```

Figure 28: Spam & ham text count

Plot the spam and ham data to evaluate the spam email messages strengths as follows:

```
In [13]: msg_df["label"].value_counts().plot(kind = 'pie', explode = [0, 0.1], figsize = (6, 6), autopct = '%1.2f%%')
plt.ylabel("Spam vs Ham")
plt.legend(["Ham", "Spam"])
plt.show()
```

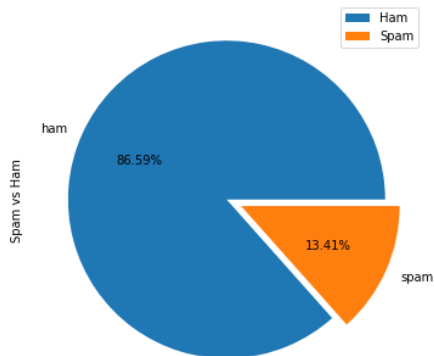


Figure 29: Ham and Spam dataset plot & check the length of text

To detect and verify the text length and shapes of the text

```
In [14]: msg_df.groupby("message")["label"].agg([len, np.max]).sort_values(by = "len", ascending = False).head(n = 10)
```

```
Out[14]:
```

	len	amax
Sorry, I'll call later	30	ham
I cant pick the phone right now. Pls send a message	12	ham
Ok...	10	ham
Ok	4	ham
Okie	4	ham
7 wonders in My WORLD 7th You 6th Ur style 5th Ur smile 4th Ur Personality 3rd Ur Nature 2nd Ur SMS and 1st "Ur Lovely Friendship"... good morning dear	4	ham
Wen ur lovable bcums angry wid u, dnt take it seriously.. Coz being angry is d most childish n true way of showing deep affection, care n luv!.. kettoda manda... Have nice day da.	4	ham
Your opinion about me? 1. Over 2. Jada 3. Kusruthi 4. Lovable 5. Silent 6. Spl character 7. Not matured 8. Stylish 9. Simple Pls reply..	4	ham
Please call our customer service representative on FREEPHONE 0808 145 4742 between 9am-11pm as you have WON a guaranteed £1000 cash or £5000 prize!	4	spam
Ok.	4	ham

Figure 30: Describing the length of the messages either its spam or ham message

```
In [15]: msg_df['length'] = msg_df['message'].apply(len)
msg_df.head()
```

```
Out[15]:
```

	label	message	length
0	ham	Go until jurong point, crazy.. Available only ...	111
1	ham	Ok lar... Joking wif u oni...	29
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155
3	ham	U dun say so early hor... U c already then say...	49
4	ham	Nah I don't think he goes to usf, he lives aro...	61

Figure 31: message length

```
In [16]: msg_df.length.describe()
```

```
Out[16]: count    5572.000000
mean         80.489950
std          59.942907
min           2.000000
25%          36.000000
50%          62.000000
75%         122.000000
max          910.000000
Name: length, dtype: float64
```

Figure 32: Statistical measurements of the text

Plot the spam and ham text as follows:

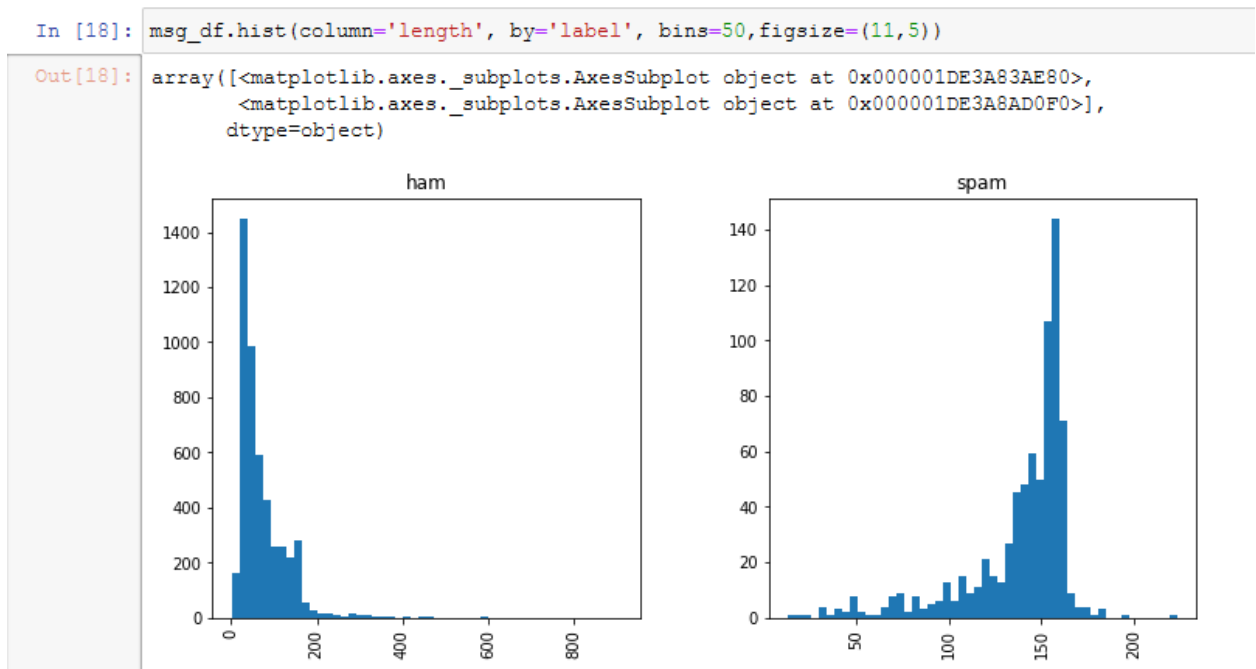


Figure 33: Plot the dataset to validation and verification of spam & ham emails

Data Cleaning & Text Transformation

In data cleaning process using the python jupyter notebook the process remove the unnecessary values from the dataset. It also remove unimportant words, stopwords and stemming

```
In [20]: import string
def cleanText(message):
    #message = message.translate(str.maketrans('ranjan', 'ranjan', string.punctuation))
    message = re.sub('[^a-zA-Z]', ' ', message)
    message = message.lower()
    message = message.split()
    words = [ss.stem(word) for word in message if word not in stop]
    return " ".join(words)

msg_df["message"] = msg_df["message"].apply(cleanText)
msg_df.head(n = 10)
```

```
Out[20]:
```

	label	message	length
0	ham	go jurong point crazi avail bugi n great world...	111
1	ham	ok lar joke wif u oni	29
2	spam	free entri wkli comp win fa cup final tkts st ...	155
3	ham	u dun say earli hor u c alreadi say	49

Figure 34: text string transformation process

NLTK library to extract spam words and ham words

The natural language processing toolkit effectively works in python due to extensive and large applications of computer vision and text transformation detection.

```
[nltk_data] Downloading package punkt to
[nltk_data] C:\Users\Ranjan\AppData\Roaming\nltk_data...
[nltk_data] Unzipping tokenizers\punkt.zip.
Out[26]: True

In [27]: spam_words = []
ham_words = []

def extractSpamWords(spamMessages):
    global spam_words
    words = [word for word in word_tokenize(spamMessages)]
    spam_words = spam_words + words

def extractHamWords(hamMessages):
    global ham_words
    words = [word for word in word_tokenize(hamMessages)]
    ham_words = ham_words + words

spam_messages.apply(extractSpamWords)
ham_messages.apply(extractHamWords)

Out[27]: 0      None
         1      None
         3      None
         4      None
         6      None
```

Figure 35: extract spam words

Finally counting the spam and ham messages to cover the various text transformation length and objects of the dataset that present the spam and ham messages detail over the period of stop words and count words after cleaning the dataset.

In [31]:

msg_df

Out[31]:

	label	message	length
0	ham	go jurong point crazi avail bugi n great world...	111
1	ham	ok lar joke wif u oni	29
2	spam	free entri wkli comp win fa cup final tkts st ...	155
3	ham	u dun say earli hor u c already say	49
4	ham	nah think goe usf live around though	61
...
5567	spam	nd time tri contact u u pound prize claim easi...	160
5568	ham	b go esplanad fr home	36
5569	ham	piti mood suggest	57
5570	ham	guy bitch act like interest buy someth els nex...	125
5571	ham	rofl true name	26

5572 rows × 3 columns

Figure 36: detecting spam and ham messages length & labels

TFIDF-IDF Vectorizer

The term frequency appears in document to specify the text on based on word counter and text presentation, this algorithm transform the text & evaluate effectively $tf(t)$, and inverse document frequency is consider as, the idf values normalize the text either it evaluate to 0 and 1 to consider the numeric text and perform various calculations.

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

Figure 37: log theorem

Transform the string text into numeric array

```
In [37]: X = cv.fit_transform(msg_df["message"]).toarray()
X
Out[37]: array([[0, 0, 0, ..., 0, 0, 0],
               [0, 0, 0, ..., 0, 0, 0],
               [0, 0, 0, ..., 0, 0, 0],
               ...,
               [0, 0, 0, ..., 0, 0, 0],
               [0, 0, 0, ..., 0, 0, 0],
               [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

```
In [38]: df = pd.DataFrame(X, columns=cv.get_feature_names())
df
df['len'] = msg_df['length']
df
Out[38]:
```

	aa	aah	aany	aaooooo	right	aathi	ab	abbey	abdomen	abeg	abel	...	zf	zhong	zindgi	zoe	zogtorius	zoom	zouk	zs	zyada	len
0	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	111
1	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	29
2	0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	155

Figure 38: transform the string text into the numeric array conversion

Train & test the data validation

Training and testing the dataset to divides the each segments values on particular columns values and submit them, training the model for using the naïve bays classifier models and bring the outcomes of spam email detections and ham emails detections after classify the models.

```
In [43]: from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df, y, test_size = 0.20, random_state = 0)

# Training model using Naive bayes classifier

from sklearn.naive_bayes import MultinomialNB
spam_detect_model = MultinomialNB().fit(X_train, y_train)

y_pred = spam_detect_model.predict(X_test)
```

Figure 39: training & testing the models

Outcomes & Results

```
In [44]: print(accuracy_score(y_test,y_pred))
print(fbeta_score(y_test,y_pred,beta =0.5))

0.9811659192825112
0.9390862944162438

In [45]: y_pred
Out[45]: array([0, 0, 0, ..., 0, 1, 0], dtype=int64)

In [46]: print (classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	0.99	0.99	0.99	955
1	0.94	0.93	0.93	160
accuracy			0.98	1115
macro avg	0.97	0.96	0.96	1115
weighted avg	0.98	0.98	0.98	1115

Figure 40: outcomes of algorithm

The above mention spam detection models present the 98% accuracy of the spam emails detections models that used to classify the each segments of text with f1 score and precision recall.

Analysis

(Bjagrav & Teja, 2020) Using natural language processing and Bayesian model to predict the text after transformation of text conversion using the method of TFIDF Vectorizer algorithm, developed three method to test & predict the method of synonym replacement in between the text modellings, technique of ham word injection and spam word spacing effectively called the machine learning models.

Chapter # 5 Discussion & Conclusion

(Zamir & Hikmat Ullah Khan, 2020) Email spam detection method using the diverse machine learning approach to detect spam and ham emails using the email dataset and spam email dataset.

Email widely used at enterprise and corporate business level that support the organization multi chain domains. So far the business center facing lot of problems and hurdles to clean the spams and malware emails. Malware email attack is minimize so the spam junction is not eradicate yet. Email server support enterprise email login and email activity facilities it also used to manage the email storage access and allocate the user domain panels.

Discoveries

Examination and investigational results used various machine learning method such as naïve bays classifier used multinomial algorithm to classify spam and ham emails using the dataset, it achieve the 98% accuracy. Beside this support vector machine and logistic regression effectively deploy with TFIDF algorithm to transform the text into numeric array and evaluate the each array cells in terms to investigate the algorithm. Streamlit is also deploy and tested in web based graphical application to detect spam and ham text, background the naïve bays algorithm has been developed that classify and integrate the text into the numeric form and predict the text according to the label based approach. Using these proposed model deep neural networks with sentiment features performed the machine learning classifier in terms of classification accuracy about 97% to 98% of detection.

Novelty

The research is novel research because streamlit concept has been introduced with the help of naïve bays algorithm. Multinomial feature extraction and feature detection algorithm deployed with the help of sklearn library using the python anaconda jupyter notebook.

(Bibi, et al., 2020) With evaluation of various machine learning algorithm on spam email detections outcomes present that the naïve bays algorithm present effective accuracy and precision using the python programming also SVM algorithm present the great score to detect the spam emails by using the dataset.

The term frequency inverse document used effectively to transform the string text into the numeric text conversion to evaluate the spam and ham text by 1 and 0 in terms of effective statistical analysis. TFIDF algorithm also used to count the word and also complement the words in the document its weights the count features.

Naïve bays classifier predict on probabilistic based which is good for text evaluation. Naïve bays called the occurrence of one feature independent of occurrences of other feature, supervised machine learning algorithm to classify the high dimensional dataset. It works on bays theorem.

$$P(A | B) = \frac{P(A \cap B)}{P(B)} \Rightarrow P(A | B).P(B) = P(A \cap B)$$

$$P(B | A) = \frac{P(B \cap A)}{P(A)} \Rightarrow P(B | A).P(A) = P(B \cap A)$$

Figure 41: probability theorem

Naïve bays calculate the conditional probability beside this on the principle of random variable.

Email Classify to Detect Spams the following steps:

1. Training
2. Extract email text
3. Parse each email token
4. Training and testing the dataset with sklearn feature extraction library
5. Filtering the dataset
6. Classify the text based on label

The performance metrics of accuracy evaluated the computed number of digits that process and present the text.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Figure 42: Accuracy pattern

(M, et al., 2020,) Logistics regression works with NLP toolkit natural language processing that process the text with countvectorizer to evaluate and integrate the dataset. Stop word and word count clean the text and clean the white spaces in the text, to measure the text performance it uses the TF-IDF algorithm that convert the text and evaluate them.

(Sharmaa & Yadav, 2021) In these days the spam mails objective has been changed now it's part of changing the scenario's , spam is used to hack the confidentially data & information, it also uses as play card to capture the corporate secrete asset information and destroy the computer networks. The objective has been changed and cyber security trying to resolve this kind of issues with development secure socket layer to hide the confidential data. Gmail services support end level security to the user, also encrypt the user data by sending to Google drives storage location. The largest file easily sent to Google drive with private and public access. Gmail services is much better than comparatively analyze to Yahoo, Outlook server. Beside this decision tree, k-nearest neighbors achieve 90% accuracy to detect spam and ham emails, social media fake profile detection decision tree algorithms helps a lot to sort out the fake emails profile id.

Future Research Challenges

(Rapacz & Chołda, 2021) Machine learning spam classifier uses the python anaconda jupyter notebook IDE platform, and it was very fastest and easy method to detect the spam and ham emails in the dataset. Data cleaning and data validation process has been carried out to find out any missing & NA values in the dataset. The approach for training and testing the dataset which is very helpful to divide the spam dataset and evaluate them to figure out the progression of spam emails and ham emails. Naïve bays classifier logistic regression classifier algorithm works with python Sklearn library that extract the feature and select the particular features validate the dataset and presents the output classifying them in appropriate manner. Comparatively naïve bays classifier works very well to detect the spam and junk dataset. Future research required to improve the yahoo and outlook emails server, comparatively Gmail server uses machine learning algorithm and works effectively and better.

Conclusion

The spam and ham email feature extraction in email dataset, python machine learning algorithms effectively works to meet the challenges and provides the interface to detect the spam emails spam id and spam text in the dataset. Dataset has been downloaded from Github and kaggle repository. Naïve bays classifier algorithm detection frequency is much better than as compare to support vector machine and logistics regression algorithm. Beside this python streamlit web based library to creative & build the spam application and shares with colleagues easily. Support vector machine classifier uses the NLTP, natural language processing mechanism to process the text in terms numeric. Logistics regression works with TFIDF classifier that transform the text into the numeric array and evaluate each bit in term of true and false. This project is going to help other researcher to clearly understand the objectives of email spam classifier and uses this project at corporate level to handle organization emails. Future research required to improve the yahoo and outlook emails server, comparatively Gmail server uses machine learning algorithm and works effectively and better.

References

- Aakash Atul Alurkar, S. B. R., 2019. A Comparative Analysis and Discussion of Email Spam Classification Methods Using Machine Learning Techniques. *Applied Machine Learning for smart data analysis*, Volume 1, pp. 1-22.
- Akinyelu, A. A., 2021. Advances in spam detection for email spam, web spam, social network spam, and review spam: ML-based and nature-inspired-based techniques. *Journal of Computer Security*, , 29 (5), pp. 473-529, .
- Alzahrani, A. & Rawat, D. B., 2019, . Comparative Study of Machine Learning Algorithms for SMS Spam Detection. *2019 SoutheastCon*,, pp. 1-6.
- Anon., , 2017, . SMS Spam Detection using H2O Framework. *Procedia Computer Science*, Volume 113, pp. 154-161.
- Bandaya, M. T. & A.Mi, F., 2011. Analyzing Internet e-mail date-spoofing. *Digital Investigation*, 7(3), pp. 145-153.
- Bibi, A., Rasia & Samina, 2020. Spam Mail Scanning Using Machine Learning Algorithm. *Journal of Computer*, 1(1), p. 1.
- Bjagrav & Teja, R., 2020 . Adversarial machine learning for spam filters 38. *ARES '20: Proceedings of the 15th International Conference on Availability, Reliability and Security*, p. 1–6 .
- CHARAN, M. C. S., 2021. MAIL CLASSIFICATION USING MACHINE LEARNING. *Journal of Engineering Sciences*, 12(8), pp. 120-130.
- Classification, E. S., 2021. *Email Spam Classification*. [Online]
Available at: <https://www.youtube.com>
[Accessed 22 November 2021].
- Clayton, R., 2007. Email Traffic: A Quantitative Snapshot. *European Journals of Investigation*, 11(1), pp. 1-10.
- Dada, E. G., 2018. Logistic Model Tree Induction Machine Learning Technique for Email Spam. *The Pacific Journal of Science and Technology* , 19(2), pp. 96-100.
- Dada, E. G. & Joseph, S. B., 2018. Random Forests Machine Learning Technique for Email Spam Filtering. *University of Maiduguri Faculty of Engineering Seminar Series*, 9(1), pp. 29-35.
- Dalkılıç, G. & Sipahi, D., 2017. Spam filtering with sender authentication network. *Computer Communications*, Volume 98, pp. 72-79.
- Dataset, S. M., 2021. *Spam Mails Dataset*. [Online]
Available at: <https://www.kaggle.com/venky73/spam-mails-dataset>
[Accessed 20 November 2021].

detection, C. m. s., 2021. *CNN method spam detection*. [Online]
Available at: www.youtube.com
[Accessed 20 November 2021].

detection, N. b. e. s., 2021. *Naive bays email spam detection*. [Online]
Available at: <https://www.youtube.com>
[Accessed 20 November 2021].

Douzi, S., AlShahwan, F. A. & Mouad., 2020. Hybrid Email Spam Detection Model Using Artificial Intelligence. *International Journal of Machine Learning and Computing*, 10(2), pp. 316-327.

Ezpeleta, E. & Mendizabal, I. V. d., 2020. Novel email spam detection method using sentiment analysis and personality recognition. *Logic Journal of the IGPL* , 28(1), p. 83–94.

Freydenberg, M. & Kevin, 2020. From Engagement to Empowermint Project based learning in python coding course. *Computer Information Systems Journal Articles*, 1(1), p. 1.

G. Chetty, H. B. a. M. W., 2019. Deep Learning Based Spam Detection System," 2019. *International Conference on Machine Learning and Data Engineering (iCMLDE)*, 2019,, pp. 91-96.

Gangavarapu, Tushaar & Jaidhar, 2020. Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artificial Intelligence Review* , 53(7), pp. 5019-5081.

Gbenga, E., Joseph, D. & Bassia, S., 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Science Direct: Heliyon Journal*, 5(6), pp. 1-50.

Gbenga, E. & StephenBassi, D. J., 2019, . Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon Science Direct*, 5,(6), pp. 1-10.

Github, 2021. *Email Spam Classifier*. [Online]
Available at: https://github.com/Chando0185/Emial_Spam_Classification
[Accessed 22 November 2021].

Govil, N., Agarwal, K. & Varshney, A. B. a. A., 2020. A Machine Learning based Spam Detection Mechanism. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, , pp. 954-957,.

Hanif Bhuiyan, A. A. T. I. J., 2018. A Survey of Existing E-mail Spam Filtering Methods Considering Machine Learning Techniques. *lobal Journal of Computer Science and Technology*,, p. 1.

Hirano, M. & Kobayashi, R., 2019. Machine Learning Based Ransomware Detection Using Storage Access Patterns Obtained From Live-forensic Hypervisor. *2019 Sixth International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, , pp. 1-6.

Hu, H., Wang, G. & Tech, V., 2021. End-to-End Measurements of Email Spoofing Attacks. *27th Usenix Symposium*, 1(1), p. 1.

- Islam, T. & Ahmed, S. L. a. N., 2019. Using Social Networks to Detect Malicious Bangla Text Content. *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, , pp. 1-4.
- Jain, A. & Bhargava, N., 2021. SPAM Filtering Using Artificial Intelligence. *Wiley online library*, Volume 1, p. 1.
- Jun, L. G., Nazir, S. & Khan, H. U., 2020. Spam Detection Approach for Secure Mobile Message Communication Using Machine Learning Algorithms. *Machine Learning and Applied Cryptography Security and communication networks*, Volume 1, p. 1.
- Kumar, A. D. ; . K. A., 2021 . A New Malware Detection Model using Emerging Machine Learning Algorithms. *International Journal of Electronics and Information Engineering* , 3(1), pp. 24 - 32 .
- Kumar, A. D., R, V. & KP, S., 2018. DeepImageSpam: Deep Learning based Image Spam Detection. *Computer Vision and Pattern Recognition (cs.CV); Cryptography and Security (cs.CR)*, p. 1.
- Kumar, K. A. a. T., 2018,. Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization,. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, , pp. . 685-690.
- M., R., Chowdary & Harith, N., 2016. EMAIL SPAM DETECTION USING PYTHON & MACHINE LEARNING. *Turkish Journal of Physiotherapy and Rehabilitation*, 3(1), pp. 1-4.
- Mail, S. S., 2021. *SVM SPAM Mail*. [Online]
Available at: [youtube.com](https://www.youtube.com)
[Accessed 20 November 2021].
- Maqsood, S. S. A. a. J., 2018. Net library for SMS spam detection using machine learning: A cross platform solution,. *2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*,, pp. 470-476.
- message, T. s. o. t. f. e. s., 2021. *The story of the first email spam message*. [Online]
Available at: <https://www.marketplace.org/2013/05/03/story-first-email-spam-message/>
[Accessed 20 November 2021].
- M, T. M., K. Y. & Thida, A., 2020,. A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification,. *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, , pp. 324-326, .
- N. Kumar, S. S. a. N., 2020,. Email Spam Detection Using Machine Learning Algorithms. *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, , pp. 108-113.
- N. L. Octaviani, E. H. R. C. A. S. a. I. M. S. D. R., 2020,. Comparison of Multinomial Naïve Bayes Classifier, Support Vector Machine, and Recurrent Neural Network to Classify Email

Spams,. 2020 *International Seminar on Application for Technology of Information and Communication (iSemantic)*, , pp. 17-21.

Nance, J. & Baumgartner, P., 2021. gobbli: A uniform interface to deep learning for text in Python. *The Journal of Open source software*, 6(62), pp. 1-10.

Neha., S., 2020. *A Study of Machine Learning Algorithms on Email Spam Classification*, Missouri State University: Southeast Missouri State University. ProQuest Dissertations Publishing, 2020. 27993789..

NurAmir, N., SjarifNurulhuda & Azmi, F., 2019, . SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm. *Procedia Computer Science*, Volume 161,, pp. 509-515.

Onashoga, A. S., O, O. & Adesina, 2015. An Adaptive and Collaborative Server-Side SMS Spam Filtering Scheme Using Artificial Immune System. *information security journals global perspective*, 24(6), pp. 1-24.

python, E. S. S. &., 2021. *EMAIL SPAM Streamlit & python detection machine learning*. [Online]
Available at: www.youtube.com
[Accessed 20 November 2021].

python, N. L. r. E. s. d., 2021. *NLP Logistics regression Email spam detection python*. [Online]
Available at: www.youtube.com
[Accessed 20 November 2021].

Raad, M., Yeassen, N. M. & Alam, G. M., 2010. Impact of spam advertisement through e-mail: A study to assess the influence of the anti-spam on the e-mail marketing. *Academic Journals Information System Department, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia*, Volume 1, pp. 1-10.

Rapacz, S. & Chołda, P., 2021. A Method for Fast Selection of Machine-Learning Classifiers for Spam Filtering. *Electronics; Basel*, 10, .(17,), p. 2083..

Roy, S. S., Sinha, A. & Roy, R., 2017. Spam Email Detection Using Deep Support Vector Machine, Support Vector Machine and Artificial Neural Network. *International Workshop Soft Computing Applications*, pp. 162-174.

Sagar, R. & Rutvij, 2020. Applications in Security and Evasions in Machine Learning: A Survey. *Electronics Journal*, 9(1), pp. 1-97.

Sahni, R., 2021. Analysis of Naive Bayes Algorithm for Email Spam Filtering. *IJMTST - International Journal for Modern Trends in Science and Technology (ISSN:2455-3778)*, 7(1), pp. 5-9.

Santoso, B., 2019. An Analysis of Spam Email Detection Performance Assessment Using Machine Learning. *Jurnal Online Informatika*, 4(1), p. 1.

- Sethi Sumesha, M., Vinayak, C. & Dahiya, C., 2021. Spam Email Detection Using Machine Learning and Neural Networks. *Sentimental Analysis and Deep Learning*, Volume 1408, pp. 275-290.
- sewayaa, Y. K., ovab, E. A. & monovb, A. A., 2021, . Evaluating the Effectiveness of Machine Learning Methods for Spam Detection. *Procedia Computer Science*, Volume 190, , pp. 479-486.
- Shahi, T. B. & Yadav, A., 2014. Mobile SMS Spam Filtering for Nepali Text Using Naïve Bayesian and Support Vector Machine. *International Journal of Intelligence Science*, , Volume 4, pp. , 24-28 .
- Sharif, B. K., bbaHyoung, A. & Kim, s., 2020. DeepCapture: Image Spam Detection Using Deep Learning and Data Augmentation. *Australasian Conference on Information Security and Privacy*, pp. 461-475.
- Sharmaa, V. D. & Yadav, S. K., 2021. An effective approach to protect social media account from spam mail – A machine learning approach. *An effective approach to protect social media account from spam mail – A machine learning approach*, Volume 1, p. 1.
- Shirani-Mehr, H. & D. Delvia Arifin, S. a. M. A. B., 2016. Enhancing spam detection on mobile phone Short Message Service (SMS) performance using FP-growth and Naive Bayes Classifier,. *2016 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, 2016,, pp. 80-84.
- Singh, A. P., Singh, A. & Chatterjee, K., 2021. A Comparative Approach for Email Spam Detection Using Deep Learning. *Intelligent Computing and Communication Systems* , pp. 187-200.
- sklearn.feature_extraction.text.CountVectorizer, 2021.
 sklearn.feature_extraction.text.CountVectorizer. [Online]
 Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
 [Accessed 22 November 2021].
- spam?, W. i. e., 2021. *What is email spam?*. [Online]
 Available at: <https://www.techtarget.com/searchsecurity/definition/spam>
 [Accessed 20 November 2021].
- Tazmina, Fobia & Kateriana, 2020. Convolutional neural networks for image spam detection. *Information Security Journal: A Global Perspective* , Volume 3, pp. 1-10.
- W. Peng, L. H. J. J. a. E. I., 2018,. Enhancing the Naive Bayes Spam Filter Through Intelligent Text Modification Detection. *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, , pp. 849-854.
- Wang, Y., 2012. Optimal Design of Hierarchical Spam Filtering Method Based on Greylisting. *Applied Mechanics & Material*, Volume 1, p. 1.

Wu, T., Liu, S. & Zhang, J., 2017. Twitter spam detection based on deep learning. *ACSW '17: Proceedings of the Australasian Computer Science Week Multiconference January 2017* , Volume , pp. 1-3.

Yang, L., Dumais, S. T. & N, P., 2017 . Characterizing and Predicting Enterprise Email Reply Behavior. *SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* , p. 235–244.

Zamir, A. & Hikmat Ullah Khan, 2020. A feature-centric spam email detection model using diverse supervised machine learning algorithms. *Emerald insights*, 38(3), p. 1.

Reflection

This research work is designed to analyze the machine learning and deep learning algorithm, this research has been used to explore the various machine learning python libraries which is beneficial for other researcher. Python IDE anaconda jupyter notebook designed particularly to develop the essentials machine learning algorithm that effectively used the machine learning strategies to use the various algorithm such as Naïve bays, support vector machine, multinomial naïve bays, logistics regression, and natural language processing tool kit, TFIDF Vectorizer and streamlit application builder. Beside this python Pycharm 2021 version has been used to design the streamlit application web based to test and verify the spam and ham emails dataset. Dataset has been purely a prime data which is part of Github repository.

Research work processes

The process of this research very clear, findings the secondary research data which is already published materials, but the factors of the finding is to verify the peer review journals including good impact factors such as ACM digital library, IEEE XPLOR, IEEE conference paper, IEEE Journals of computer science. Science Direct. MDPI research. Beside this Google scholars has been used to collect the research finding. The research methodology idea has been taken from secondary data which essentially deploy the various machine learning algorithm such as Naïve bays, SVM, LG, NLP, streamlit, Multinomial Naïve bays. TFIDF Vectorizer. But the issue exist in previous research that does not combine all of these algorithm into one research. So this research is carrying a novel approach to combine these approach in one research.

Research idea

The research methodology idea has been taken from the secondary data which has been develop the various spam email detection algorithm effectively.

Skills development

During this research project, I have learned a lot about research methods, various machine learning technology algorithms that help me a lot in my future research work. I have developed various skills during my research project that help other researchers to design machine learning algorithms. I have learned a lot from YouTube video lectures and learned from Github repositories that was a very amazing experience.

Teacher meeting

My supervisor and my teammates help me a lot during this project. My supervisor essentially supports me in every step of developing this unique work. It supports me that how I learn machine learning and deep learning algorithms to get the work more beneficial for other researchers.

Research experience

The research experience was very amazing, streamlit email spam detector that is very amazing python technology that builds the application quickly to analyze the spam and ham emails with the help of multinomial naïve bayes algorithm. The pattern of methodology I learn from various YouTube tutorials and Google search engine helps me in this regard. My friends and my parents and my university supervisor are the blessings of Almighty great Allah. I am really thankful all of you in this regard. I dedicate this project to my lovely parents. This research project is designed to combine all of the machine learning algorithms into one platform, future research required to improve the yahoo, outlook emails platform. Gmail works very well because of development of machine learning algorithms.