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

Chapter # 3: Methodology

Student Name: Shoaib

Table of Contents

[Chapter # 3 Methodology 3](#_Toc95836697)

[Introduction 3](#_Toc95836698)

[Research Philosophy 3](#_Toc95836699)

[Data Collection and interpretation 4](#_Toc95836700)

[Research Strategy 4](#_Toc95836701)

[Gmail Spam Filter: 4](#_Toc95836702)

[Yahoo Mail Spam Filters: 5](#_Toc95836703)

[Outlook Email Spam Filter: 5](#_Toc95836704)

[Data Collection Method 6](#_Toc95836705)

[Secondary Dataset: 6](#_Toc95836706)

[Secondary Data for Research finding: 6](#_Toc95836707)

[Data Analysis Method 6](#_Toc95836708)

[Naïve bays classifier: 6](#_Toc95836709)

[SVM support vector machine classifier 7](#_Toc95836710)

[Natural Language processing using python with Logistics regression spam email classifier 7](#_Toc95836711)

[Convolutional Neural Network deep learning algorithm 8](#_Toc95836712)

[Email Spam Classification App that build by using the streamlit & python 8](#_Toc95836713)

[Quantitative Method 8](#_Toc95836714)

[ Statistical machine learning method 8](#_Toc95836715)

[ GUI application 8](#_Toc95836716)

[Results of Research 9](#_Toc95836717)

[Sources 10](#_Toc95836718)

[Ethical Concern 11](#_Toc95836719)

[References 12](#_Toc95836720)

# 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. Secondary dataset is collected from Kaggle and Github repository.

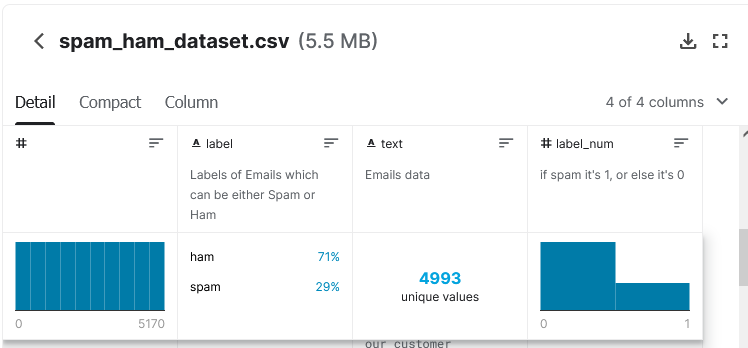
## Research Philosophy

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 secondary 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: 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

### Secondary Dataset:

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

### Secondary Data for Research finding:

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 secondary dataset 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 ntlk 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

1. 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

1. 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

1. 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

1. 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, secondary dataset 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.

# References

Clayton, R., 2007. Email Traffic: A Quantitative Snapshot. *European Journals of Investigation,* 11(1), pp. 1-10.

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].

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.

Mail, S. S., 2021. *SVM SPAM Mail.* [Online]   
Available at: youtube.com  
[Accessed 20 November 2021].

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].

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 1Information System Department, Faculty of Computer Science and Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia,* Volume 1, pp. 1-10.

spam?, W. i. e., 2021. *What is email spam?.* [Online]   
Available at: https://www.techtarget.com/searchsecurity/definition/spam  
[Accessed 20 November 2021].