

Project Synopsis  
on  
**Unsolicited email and SMS-Shot**

Submitted as a part of course curriculum for

**Bachelor of Technology**  
in  
**Computer Science**



**Submitted by**

Abhijeet Kannaujia(2000290120007)

**Under the Supervision of**  
Professor Sreesh Gaur  
Designation

**KIET Group of Institutions, Ghaziabad**  
**Department of Computer Science**  
**Dr. A.P.J. Abdul Kalam Technical University**  
**2022-2023**

## **DECLARATION**

We hereby declare that this submission is our work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

Signature of Students :

Name: Abhijeet Kannaujia

Roll No.: 2000290120007

Date:

## **CERTIFICATE**

This is to certify that Project Report entitled “**Unsolicited email and SMS-Shot**” which is submitted by **Abhijeet kannaujia** in partial fulfilment of the requirement for the award of degree B. Tech. in Department of Computer Science of Dr A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

**Date:**

**Supervisor Signature**

Supervisor Name

(Designation)

## **ACKNOWLEDGEMENT**

It gives us a great sense of pleasure to present the synopsis of the B.Tech Mini Project undertaken during B.Tech. Third Year. We owe a special debt of gratitude to **PROFESSOR SREESH GAUR**, Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his/her constant support and guidance throughout the course of our work. **His** sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his/her cognizant efforts that our endeavours have seen the light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Ajay Kumar Shrivastava, Head of the Department of Computer Science, KIET Group of Institutions, Delhi- NCR, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

Last but not the least, we acknowledge our friends for their contribution to the completion of the project.

Signature:

Date:

Name: Shubhi

Roll No: 2000290120160

Name: Abhijeet Kannaujia

Roll No.: 2000290120007

Name: Shikhar Raj

Roll No.: 2000290120144

## **ABSTRACT**

This project aims to early detection and analysis of malicious emails, SMS which can cause intrusion and unwanted access to the individual system thus compromising user data and how the Machine learning is the one of the promising answers that can be effective against this threat. In this project the naive bayes classifier, support vector, decision tree algorithms have been used. Together these algorithms combine in to perform these tasks. This project aims for an easy yet sophisticated implementation of such algorithms to design a machine that can withstand those external cyber threats and helps in neutralization of those threats.

# TABLE OF CONTENTS

|                                      | Page<br>No. |
|--------------------------------------|-------------|
| TITLE PAGE .....                     | i           |
| DECLARATION .....                    | ii          |
| CERTIFICATE .....                    | iii         |
| ACKNOWLEDGEMENT.....                 | iv          |
| ABSTRACT.....                        | v           |
| <br>                                 |             |
| CHAPTER 1 INTRODUCTION               | 1           |
| 1.1. Introduction .....              | 1           |
| 1.2 Problem Statement.....           |             |
| 1.2. Objective.....                  | 2           |
| 1.3. Scope.....                      | 3           |
| CHAPTER 2 LITERATURE REVIEW.....     | 7           |
| CHAPTER 3 PROPOSED METHODOLOGY ..... | 8           |
| 3.1 Flowchart                        | 9           |
| 3.2 Algorithm Proposed               | 10          |
| CHAPTER 4 TECHNOLOGY USED .....      | 12          |
| CHAPTER 5 DIAGRAMS .....             | 13          |
| CHAPTER 6 CONCLUSION .....           | 14          |
| REFERENCES.....                      |             |

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 BACKGROUND**

Today, spam has become a big internet issue. Recent 2017, the statistic shown spam accounted for 55% of all e-mail messages, same as during the previous year. Spam which is also known as unsolicited bulk email has led to the increasing use of email as email provides the perfect ways to send the unwanted advertisement or junk newsgroup posting at no cost for the sender. This chance has been extensively exploited by irresponsible organizations and resulting to clutter the mail boxes of millions of people all around the world. Evolving from a minor to major concern, given the high offensive content of messages, spam is a waste of time. It also consumed a lot of storage space and communication bandwidth. End user is at risk of deleting legitimate mail by mistake. Moreover, spam also impacted the economical which led some countries to adopt legislation. Text classification is used to determine the path of incoming mail/message either into inbox or straight to spam folder. It is the process of assigning categories to text according to its content. It is used to organized, structures and categorize text. It can be done either manually or automatically. Machine learning automatically classifies the text in a much faster way than manual technique. Machine learning uses pre-labelled text to learn the different associations between pieces of text and it output. It used feature extraction to transform each text to numerical representation in form of vector which represents the frequency of word in predefined dictionary. Text classification is important in the context of structuring the unstructured and messy nature of text such as documents and spam messages in a cost-effective way. A Machine learning platform has capabilities to improve the accuracy of predictions. About Big Data, a Machine Learning platform has abilities to speed up analyzing of gigantic data. It is important especially to a company to analyze text data, help inform business decisions and even automate business processes. For example, text classification is used in classifying short texts such as tweets or headlines. It can be used in larger documents such as media articles. It also can be applied to social media monitoring, brand monitoring etc. In this project, a machine learning technique is used to detect the spam message of a mail. Machine learning is where computers can learn to do something without the need to explicitly program them for the task. It uses data and produce a program to perform a task such as classification. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in machine learning studio. A specific algorithm is used to learn the classification rules from these messages. Those algorithms are used for classification of objects of different classes. The algorithms are provided with input and output data and have a self-learning program to solve the given task. Searching for the best algorithm and model can be time consuming. The two-class classifier is best used to classify the type message either spam or ham. This algorithm is used to predict the probability and classification of data outcome.

## **1.2 PROBLEM STATEMENT**

A tight competition between filtering method and spammers is going on per day, as spammers began to use tricky methods to overcome the spam filters like using random sender addresses or append random characters at the beginning or end of mails subject line. There are a lack of machine learning focuses on the model development that can predict the activity. Spam is a waste of time to the user since they must sort the unwanted junk mail and it consumed storage space and communication bandwidth. Rules in other existing must be constantly updated and maintained make it more burden to some user and it is hard to manually compare the accuracy of classified data.

## **1.3 OBJECTIVES**

**There are four objectives that need to be achieved in this project:**

- i. To study on how to use machine learning techniques for spam detection.
- ii. To modify machine learning algorithm in computer system settings.
- iii. To leverage modified machine learning algorithm in knowledge analysis software.
- iv. To test the machine learning algorithm real data from machine learning data repository.

## **1.4 PROJECT SCOPE**

This project needs a coordinated scope of work. These scopes will help to focus on this project. The scopes are:

- i. Modified existing machine learning algorithm.
- ii. Make use and classify of a data set including data preparation, classification and visualization.
- iii. Score of data to determine the accuracy of spam detection

## **1.5 LIMITATION OF WORK**

- i. This project can only detect and calculate the accuracy of spam messages
- ii. It focusses on filtering, analyzing and classifying the messages.
- iii. Do not block the messages.



## **CHAPTER-2**

### **LITERATURE REVIEW**

#### **1: Email Spam Detection Using Machine Learning Algorithms**

Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious link through spam emails which can harm our system and can also seek in into your system. Creating a fake profile and email account is much easy for the spammers, they pretend like a genuine person in their spam emails, these spammers target those peoples who are not aware about these frauds. So, it is needed to Identify those spam mails which are fraud, this project will identify those spam by using techniques of machine learning, this paper will discuss the machine learning algorithms and apply all these algorithms on our data sets and best algorithm is selected for the email spam detection having best precision and accuracy.

#### **2: A Machine Learning based Spam Detection Mechanism**

In today's internet-oriented data; receiving spam email messages are quite obvious. Most of the time such emails are commercial. But many times, such emails may contain some phishing links that have malware. This arises the need for proposing prudent mechanism to detect or identify such spam emails so that time and memory space of the system can be saved up to a great extent. In this paper, we presented the same mechanism which can filter spam and non-spam emails. Our proposed algorithm generates dictionary and features and trains them through machine learning for effective results.

#### **3: Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization**

Now-a-days, communication through email has become one of the cheapest and easy ways for the official and business users due to easy availability of internet access. Most of the people prefer to use email to share important information and to maintain their official records. But just like the two sides of coin, many people misuse this easy way of communication by sending unwanted & useless bulk emails to others. These unwanted emails are spam emails that affect the normal user to face the problems like excessive usage of their mailbox memory and filtration of useful email from unwanted useless emails. So, there is the need of some autonomous approach that filters the excessive data of emails in the form of spam emails. In this paper, an integrated approach of machine learning based Naïve Bayes (NB) algorithm and computational intelligence-based Particle Swarm Optimization (PSO) is used for the email spam detection. Here, Naïve Bayes algorithm is used for the learning and classification of email content as spam and non-spam. PSO has the stochastic distribution & swarm behaviour. property and considered for the global optimization of the parameters of NB approach. For experimentation, dataset of Ling spam dataset is considered and evaluated the performance in terms of precision, recall, f-measure and accuracy. Based on the evaluated results, PSO outperforms in comparison with individual NB approach.

#### **4: A Review on Machine Learning Techniques for Image Based Spam Emails Detection**

Sending and receiving e-mails have continued to take the lead being the easiest and fastest way of communication despite the presence of other forms of communication such as social networking. The rise in online transactions through email has globally contributed to the increasing rate of spam emails relatively which has been a major problem in the field of computing. In this note, there are many machine learning techniques available for detecting these unwanted spams. In spite of the significant progress made in the figures of literature reviewed, there is no machine learning method that has achieved 100% accuracy. Each algorithm only utilizes limited features and properties for classification. Therefore, identifying the best algorithm is an important task as their strengths need to be weighed against their limitations. In this paper we explored different machine learning techniques relevant to the spam detection and discussed the contributions provided by researchers for controlling the spamming problem using machine learning classifiers by conducting a comparative study of the selected machine learning algorithms such as: Naive Bayes, Clustering techniques, Random Forest, Decision Tree and Support Vector Machine (SVM).

## **5: A study of machine learning classifiers for spam detection**

In the present world, there is a need of emails communication but unsolicited emails hamper such communications. The present research emphasizes to build a spam classification model with/without the use of ensemble of classifiers methods have been incorporated. Through this study, the aim is to distinguish between ham emails and spam emails by making an efficient and sensitive classification model that gives good accuracy with low false positive rate. Greedy Stepwise feature search method has been incorporated for searching informative feature of the Enron email dataset. The comparison has been done among different machine learning classifiers (such as Bayesian, Naive Bayes, SVM (support vector machine), J48 (decision tree), Bayesian with Ad boost, Naive Bayes with Ad boost). The concerned classifiers are tested and evaluated on metric (such as F-measure (accuracy), False Positive Rate, and training time). By analyzing all these aspects in their entirety, it has been found that SVM is the best classifier to be used. It has the high accuracy and the low false positive rate. However, training time of SVM to build the model is high, but as the results on other parameters are positive, the time does not pose such an issue.

## **6: Machine learning in cybersecurity**

This analysis is made with the use of a contaminated data sets, and python tools for developing machine learning for detect phishing attacks through of the analysis of URLs to determinate if it is good or bad URLs in base of specific characteristics of the URLs, with the goal of provide Realtime information for take proactive decisions that minimize the impact of an attack. Machine learning plays an essential role in threat detection and prevention if integrated with the cybersecurity. For instance, according to Google, 50-70% of emails processed through their Gmail client are spam Using ML algorithms, Google is making it possible to block such unwanted content with 99% accuracy. The proposed framework successfully extracts the features from the web pages and performs a successful detection process for the phishing attack. In the multi-layered feed-forwarding framework, the first layer utilizes Random Forest, Support Vector Machine, and K-Nearest Neighbor classifiers to build a model for detecting malware from the real-time input.

## **7: Detecting Port Scan Attempts with Comparative Analysis of Deep Learning and Support Vector Machine Algorithms**

The usage of new technologies provides great benefits to individuals, companies, and governments, however, it uses some problems against them. For example, the privacy of important information, security of stored data platforms, availability of knowledge etc. Depending on these problems, cyber-terrorism is one of the most important issues in today's world. The performance measurements of support vector machine and deep learning algorithms based on up-to-date CICIDS2017 dataset were presented comparatively. Results show that the deep learning algorithm performed significantly better results than SVM. The CICIDS2017 dataset is used in our study. The dataset is developed by the Canadian Institute for Cyber Security and includes various common attack types.

## **8: Trojan Traffic Detection Based on Machine**

This analysis is based on the network behavior features and network traffic of several typical Trojans such as Zeus and Weasel, and proposes a Trojan traffic detection algorithm based on machine learning. First, model different machine learning algorithms and use Random Forest algorithm to extract features for Trojan behavior and communication features. First, we need to capture Trojan traffic, distinguish the known Trojan traffic data packets from normal traffic data packets, and then match them in a certain proportion to form the final sample data set for machine learning algorithm model training. Part of the data extracted from the training data set is used as the test set to perform performance detection and evaluation on related algorithm models. The experiment is conducted in a virtual machine intranet experiment environment built by a local computer and VMware, and the data set used is the IDS-2017 Trojan data set. Through the processing of the obtained data set, we selected three kinds of Trojan traffic data from various Trojan traffic data to conduct experiments, namely Weasel, Zeus and Zero Access Trojan traffic data. However, from all aspects, the evaluation accuracy and precision of the model trained by the Random Forest algorithm are higher than the other two machine learning algorithms, and it can also be seen from the evaluation of each model performed by the F1-Score value that the Random Forest algorithm Compared with the other two machine learning algorithms, the Random Forest algorithm has a better model detection effect.

## **9. A User-Centric Machine Learning Framework for Cyber Security Operations Center**

To assure cyber security of an enterprise, typically SIEM (Security Information and Event Management) system is in place to normalize security events from different preventive technologies and flag alerts. Analysts in the security operation centre (SOC) investigate the alerts to decide if it is truly malicious or not. However, generally the number of alerts is overwhelming capacity to handle all alerts. Because of this, potential malicious attacks and compromised hosts may be missed. Machine learning is a viable approach to reduce the false positive rate and improve the productivity of SOC analysts. In this analysis, we present a user-centric machine learning system which leverages big data of various security logs, alert information, and analyst insights to the identification of risky user. This system provides a complete framework and solution to risky user detection for enterprise security operation centre.

## **10: Adaptive Machine Learning: A framework on active malware detection**

Applications of Machine Learning (ML) algorithms in cybersecurity provide significant performance enhancement over traditional rule-based algorithms. These intelligent cybersecurity solutions demand careful integration of the learning algorithms to develop a significant cyber incident detection system to formulate security analysts industrial level. The development of advanced malware programs poses a critical threat to cybersecurity systems. Hence, an efficient, robust, and scalable malware recognition module is essential for every cybersecurity product. Conventional Signature-based methods struggle in terms of robustness and effectiveness during malware detection, specifically in the case of zero-day and polymorphic viruses attacks. In this paper, we design an adaptive Machine Learning based active malware detection framework which provides a cybersecurity solution against phishing attacks. The proposed framework utilize ML algorithms in a multilayered feed-forwarding approach to successfully detect the malware by examining the static features of the web pages. The proposed framework successfully extracts the features from the web pages and performs a successful detection process for the phishing attack. In the multilayered feed-forwarding framework, the first layer utilizes Random Forest (RF), Support Vector Machine (SVN), and K-Nearest Neighbor (K-NN) classifiers to build a model for detecting malware from the real-time input. The output of the first layer passes to the Ensemble Voting (EV) algorithm, which accumulates earlier classifiers performance. At the third layer, adaptive frameworks investigate second layer input data and formulate the phishing detection model. We analyze the proposed frameworks performance on three different phishing datasets and validate the higher accuracy rate. Our goal is to classify a given web page as phishing or not. The problem is formulated as a binary classification task. Consider as set of  $w$  webpages  $(u_1, y_1)(u_2, y_2), \dots, (u_w, y_w)$  where  $u_w$  for  $w = 1, 2, 3, \dots, W$  represents a Webpage and  $y_w \in [-1, 1]$  represents the label of webpage instance with  $y_w = +1$  being a Phishing Webpage and  $y_w = -1$  being a benign webpage. The classification procedure is to obtain a feature representation  $u_w \in \mathbb{R}^n$  where  $\mathbb{R}^n$  is the  $n$  dimensional feature vector representation Webpage  $u_w$ . The next step is to learn a prediction function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  which is the score predict-ing the class assignment for webpage instance  $x$ . The prediction made by the function  $f$  the can minimize the total number of mistakes  $(\sum_{w=1}^W \mathbb{I}(f(u_w) \neq y_w))$  in the entire dataset.

Our experimental assessment results are productive because of performance metrics. Our proposed Active Malware Detection is demonstrated based on inspecting our proposed algorithms with Precision, recall, and F-measure metrics. In previous works, the attacks like phishing attacks were detected using the conventual methods, which are now ineffective due to the emergence of computer technologies (i.e., Machine learning, Deep learning). The methodology, whereas we introduce a modern way to train the model for possible future attacks through feedforwarding technique, which will provide resistance to adversarial attacks on the machine learning model, one of the machine learning model problems working in a real-time environment.

## **11: AI in CyberSecurity education**

Machine learning (ML) techniques are changing both the offensive and defensive aspects of cybersecurity. The implications are especially strong for privacy, as ML approaches provide unprecedented opportunities to make use of collected data. Thus, education on cybersecurity and AI is needed. To investigate how AI and cybersecurity should be taught together, we look at previous studies on cybersecurity MOOCs by conducting a systematic literature review. The initial search resulted in 72 items and after screening for only peer-reviewed publications on cybersecurity online courses, 15 studies remained. Three of the studies concerned multiple cybersecurity MOOCs whereas 12 focused on individual courses. The number of published work evaluating specific cybersecurity MOOCs was found to be small compared to all available cybersecurity MOOCs. Analysis of the studies revealed that cybersecurity education is, in almost all cases, organized based on the topic instead of used tools, making it difficult for learners to find focused information on AI applications in cybersecurity. Furthermore, there is a gap in academic literature on how AI applications in cybersecurity should be taught in online courses. We summarize the findings from the literature review in four points: • There exists relatively few case studies on cybersecurity MOOCs compared to existing available MOOCs. • Cybersecurity MOOCs organize educational content in most cases based on covered topics instead of the methods (such as AI). • Even from the most recent MOOCs, only a few mention teaching ML techniques applied to cybersecurity. • Domain-specific pedagogical studies on how to teach AI applied in cybersecurity, or which applications of AI should be covered, are missing. There are relatively few case studies on actual cybersecurity MOOCs. Designers looking into best practices or pedagogical strategies on how to teach certain topics would arguably benefit from such studies. Furthermore, following the result that there is a lack of MOOCs on applications of AI in cybersecurity, future work could involve creating more of such MOOCs focusing on key technologies and areas where ML is relevant. Undoubtedly such courses would be on the advanced level, narrowing down the number of participants looking for them, which might discourage some parties from committing resources into creating those MOOCs. We investigated how the application of AI has been taught in cybersecurity MOOCs and what design philosophies exist by systematically reviewing existing peer-reviewed studies. The results showed that there are surprisingly few studies concerning cybersecurity MOOCs compared to the amount of courses currently offered. Furthermore, all courses, which were discussed in the papers, were organised based on their topic, and none based on the applied method (such as AI). This can be limiting for students looking to specifically learn about how AI is used in the domain of cybersecurity. Finally, only a couple of courses mentioned AI in their course content. These challenges have been addressed by previous work by suggesting that the industry would work together with academia to update course materials to include AI [13]. Altogether, the rapid increase in the popularity of ML applications does not yet show in studies on cybersecurity MOOCs. Updates on existing courses are required to ensure learners receive up to date information on the impact of AI on cybersecurity. New courses could look into organizing content based on which applications of AI are most relevant for cybersecurity.

## **12: Machine Learning and Cyber Security**

The application of machine learning (ML) technique in cybersecurity is increasing than ever before. Starting from IP traffic classification, filtering malicious traffic for intrusion detection, ML is the one of the promising answers that can be effective against zero day threats. New research is being done by use of statistical traffic characteristics and ML techniques. This paper is a focused literature survey of machine learning and its application to cyber analytics for intrusion detection, traffic classification and applications such as email filtering. Based on the relevance and the number of citation each methods were identified and summarized. Because datasets are an important part of the ML approaches some well know datasets are also mentioned. Some recommendations are also provided on when to use a given algorithm. An evaluation of four ML algorithms has been performed on MODBUS data collected from a gas pipeline. Various attacks have been classified using the ML algorithms and finally the performance of each algorithm have been assessed. In this paper an elaborate survey was performed to enlist few popular datasets then few ML algorithms were discussed along with their application in cybersecurity. Finally few recommendations were made regarding the choice of ML. In the later part of the paper a brief analysis was performed with an ICS data set and performance of a few ML algorithm was evaluated. Although J48 algorithm performs better than other algorithms in the scope of analysis, more analysis needs to be performed to ascertain the performance of the algorithms because the performance of algorithms tends to skewed depending upon the dataset on which it is being applied on. Secondly, Random forest might be more suitable as a core IDS algorithm for its optimal real-time.

## **13: ML learnings techniques for spam detection in e-mail and iot platforms**

Nowadays, emails are used in almost every field, from business to education. Emails have two subcategories, i.e., ham and spam. Email spam, also called junk emails or unwanted emails, is a type of email that can be used to harm any user by wasting his/her time, computing resources, and stealing valuable information. (e ratio of spam emails is increasing rapidly day by day. Spam detection and filtration are significant and enormous problems for email and IoT service providers nowadays. Among all the techniques developed for detecting and preventing spam, filtering email is one of the most essential and prominent approaches. Several machine learning and deep learning techniques have been used for this purpose, i.e., Naïve Bayes, decision trees, neural networks, and random forest. (is paper surveys the machine learning techniques used for spam filtering techniques used in email and IoT platforms by classifying them into suitable categories. A comprehensive comparison of these techniques is also made based on accuracy, precision, recall, etc. In the end, comprehensive insights and future research directions are also discussed. In the last two decades, spam detection and filtration gained the attention of a sizeable research community. (e reason for a lot of research in this area is its costly and massive effect in many situations like consumer behavior and fake reviews. (e survey covers various machine learning techniques and models that the various researchers have proposed to detect and filter spam in emails and IoT platforms. (e study categorized them as supervised, unsupervised, reinforcement learning, etc. (e study compares these approaches and provides a summary of learned lessons from each category. (is study concludes that most of the proposed email and IoT spam detection methods are based on supervised machine learning techniques. A labeled dataset for the supervised model training is a crucial and time-consuming task. Supervised learning algorithms SVM and Naïve Bayes outperform other models in spam detection. (e study provides comprehensive insights of these algorithms and some future research directions for email spam detection and filtering.

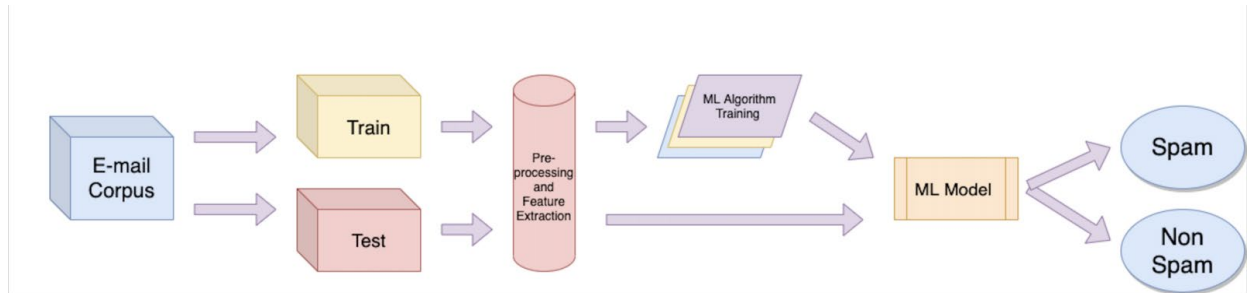
## **14: ML based detection of Spam e-mails**

Social communication has evolved, with e-mail still being one of the most common communication means, used for both formal and informal ways. With many languages being digitized for the electronic world, the use of English is still abundant. However, various native languages of different regions are emerging gradually. The Urdu language, coming from South Asia, mostly Pakistan, is also getting its pace as a medium for communications used in social media platforms, websites, and emails. With the increased usage of emails, Urdu's number and variety of spam content also increase. Spam emails are inappropriate and unwanted messages usually sent to breach security. These spam emails include phishing URLs, advertisements, commercial segments, and a large number of indiscriminate recipients. Thus, such content is always a hazard for the user, and many studies have taken place to detect such spam content. However, there is a dire need to detect spam emails, which have content written in Urdu language. The proposed study utilizes the existing machine learning algorithms including Naive Bayes, CNN, SVM, and LSTM to detect and categorize e-mail content. According to our findings, the LSTM model outperforms other models with a highest score of 98.4% accuracy.

The results and comparisons of different classifiers after data training and testing are presented in this section. We gathered 5000 emails from the online resource 'kaggle' and translated them into Urdu using the python library Google trans, which uses the Google Translate Ajax API. Four thousand emails were used to train various ML and DL models. One thousand emails were used for testing in order to quantify accuracy and assessment metrics. As explained about evaluation measures in section 5, we have evaluated accuracy, precision, recall, and f-measures that are evaluation measures measured using SVM and Naive Bayes. CNN and LSTM are used to measure ROC-AUC and model loss values. Finally, using various graphs, a comparison of models is presented below. The findings in Table 4 show that the deep learning algorithm (LSTM) is a stronger method for detecting Urdu spam emails, with high accuracy of 98.4%. With the increase usage of emails, this study focuses on using automated ways to detect spam emails written in Urdu. The study uses various machine learning and deep learning algorithms to detect them. In the study, a translated emails dataset including spam and ham emails is generated from Kaggle, which is preprocessed for various approaches. Accuracy, precision, recall, F-measure, ROC-AUC, and model loss are used as comparative measures to examine performance. The study concludes that deep learning models are more successful in classifying Urdu spam emails. Comparatively, LSTM algorithm has a high accuracy rate of around 98% with low model loss rate of 5%. Even though LSTM takes a little longer to train than CNN, SVM, or Naive Bayes, its efficiency and accuracy rate are far better than those of the other approaches. The creation of an actual dataset of Urdu emails can be considered as a viable future task. In addition, more recent artificial intelligent approaches may also be considered to detect spams.

## **CHAPTER : 3**

### **METHODOLOGY**



### **IMPLEMENTATION AND CODING PHASE**

This project is developed by using Python Language and combining with the Vow pal Wabbit algorithm. Azure machine learning studio are as the platform to develop the project. It contains important function for preprocessing the dataset. Then, the dataset is going to be used to train and test the model of the machine learning achieve the objectives of the project.

### **PROJECT REQUIREMENT AND SPECIFICATION**

System requirement is needed in order to accomplish the project goals and objectives and to assist in development of the project that involves the usage of hardware and software. Each of these requirements is related to each other to make sure that system can be done smoothly.

#### **HARDWARE**

**The usage of hardware is as below**

| No. | Hardware | Type             | Description   |
|-----|----------|------------------|---|
| 1.  | Laptop   | Acer Aspire E 14 | <ul style="list-style-type: none"><li>· Processor: Intel Core i5, 7<sup>th</sup> Gen</li><li>· OS version: Windows 64 bit</li><li>· RAM: 8 GB</li></ul> |
| \2. | Printer  | HP Deskjet 2135  | <ul style="list-style-type: none"><li>· Printing document</li></ul>   |



|    |                   |          |   |
|----|-------------------|----------|---|
| 3. | Printed paperwork | A4 paper | · Used to study on how to implement this project from pastpaperwork |
|----|-------------------|----------|---|

*Table 1: Hardware used*

*Table 3.1: Hardware used*

It is better to use high performance processor to avoid any problem while doing this project. Machine learning project required a high-speed processor for a better performance to train a large amount of data.

## **SOFTWARE**

**The usage of software in this project is as below**

| No. | Software            | Description   |
|-----|---------------------|---|
| 1.  | Microsoft Azure     | <ul style="list-style-type: none"> <li>· Machine learning platform</li> <li>· Deploy models</li> <li>· Run models in cloud</li> </ul> |
| 2.  | Google Chrome       | <ul style="list-style-type: none"> <li>· Used to run web-based system</li> </ul>  |
| 3.  | Microsoft Word 2016 | <ul style="list-style-type: none"> <li>· Creating and editing report</li> </ul>   |

|     |                                 |  |
|-----|---------------------------------|--|
| 4.  | Microsoft PowerPoint 2016       | · For presenting finding and result of the project |
| 5.  | Github                          | · Get dataset                                      |
| 6.  | Kaggle                          | · Get dataset                                      |
| 7.  | UCI machine learning repository | · Get dataset                                      |
| 8.  | Snipping Tool                   | · Captures and screenshot images                   |
| 9.  | WinZip                          | · Extract the data                                 |
| 10. | Visual Studio                   | · Implementation and deployment                    |

*Table .2: Software used*

## **FRAMEWORK**

### **DATA SOURCE**

Collecting data is utterly difficult due to numerous constraints for instances the volume of data and the throughput required for proper and timely ingestion. The dataset that I've used in this project is the real existing data that can be downloaded from machine learning data repository site. There are three websites that I've visit to get the dataset to be used in this project.



*Figure 1: Data set a*

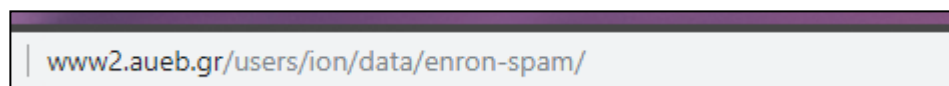


Figure 2: Data set b








Figure 3: Data set c

## DATA SETS

A figure 3.4 shows the list of data set provided by each website. These datasets might contain more than 1000 labelled messages for training and testing. The data first need to be reformatted into .CSV by splitting them into training.csv and testing.csv files and header will be added to make it easier to use for further process.




- Enron-Spam in raw form:
  - ham messages:
    - [beck-s](#)
    - [farmer-d](#)
    - [kaminski-v](#)
    - [kitchen-l](#)
    - [lokey-m](#)
    - [williams-w3](#)
  - spam messages:
    - [BG](#)
    - [GP](#)
    - [SH](#)

Figure 4: List of data set by GitHub

| Index of /ml/machine-learning-databases/spambase   |                   |      |             |
|--|-------------------|------|-------------|
| Name   | Last modified     | Size | Description |
|  <a href="#">Parent Directory</a>       |                   | -    |             |
|  <a href="#">spambase.DOCUMENTATION</a> | 20-Aug-1999 11:21 | 6.3K |             |
|  <a href="#">spambase.data</a>          | 20-Aug-1999 11:21 | 686K |             |
|  <a href="#">spambase.names</a>         | 20-Aug-1999 11:21 | 3.5K |             |
|  <a href="#">spambase.zip</a>           | 20-Aug-1999 11:21 | 123K |             |

Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 443

Figure 5: List of data set by UCI

|   |   |  |  |
|---|---|--|--|
| 0 |  | <b>spam messages</b><br>Owais updated a year ago (Version 1)   | <div> <div>CSV</div> <div>207.8 KB</div> <div>CC0</div> </div> <div> <div>&lt;/&gt; 1</div> <div>0</div> <div>605</div> </div>   |
| 1 |  | <b>Spam Identification</b><br>lianglirong updated 3 months ago (Version 1)   | <div> <div>Other</div> <div>1.2 MB</div> <div>Other</div> </div> <div> <div>&lt;/&gt; 2</div> <div>0</div> <div>136</div> </div> |
| 3 |  | <b>Spam Text Message Classification</b><br>Let's battle with annoying spammer with data science.<br>Team AI updated a year ago (Version 1) | <div> <div>CSV</div> <div>205.2 KB</div> <div>CC0</div> </div> <div> <div>&lt;/&gt; 10</div> <div>1</div> <div>2k</div> </div>   |

*Figure 6: List of data set by Kaggle*

## PROCESS MODEL

Process model is a series of steps, concise description and decisions involved in order to complete the project implementation. In order to finish the project within the time given, the flows of project need to be followed. The framework below shows how the overall flow of this project in order to separate between a spam and ham message.

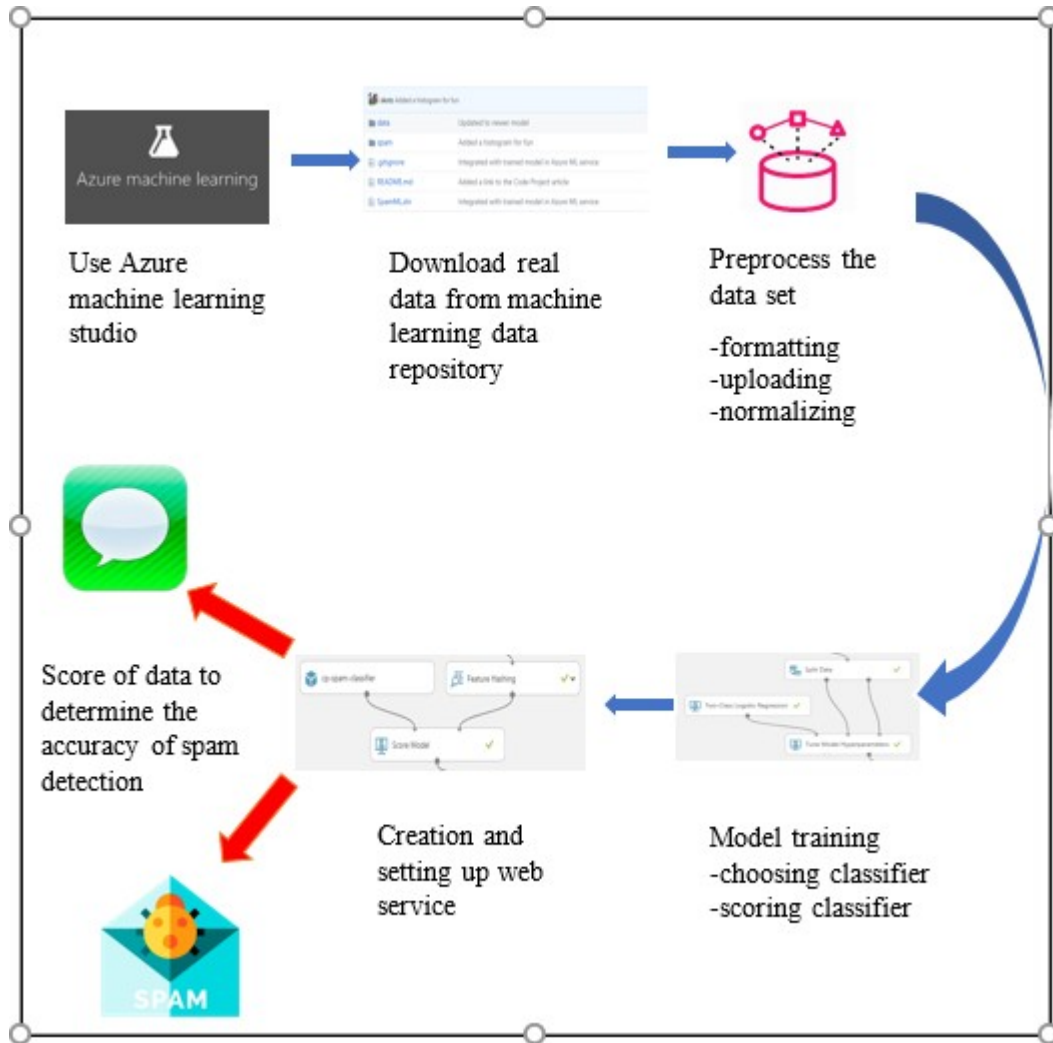


Figure 7: Process framework

## DATA MODEL

As for data model, it refers to the documenting a complex system and data flow between different data elements and design as an easily understood diagram using text and symbol. The data flow below shows how the data flow of these project in order to detect the spam messages and classify them into two separate type which is spam and ham message.

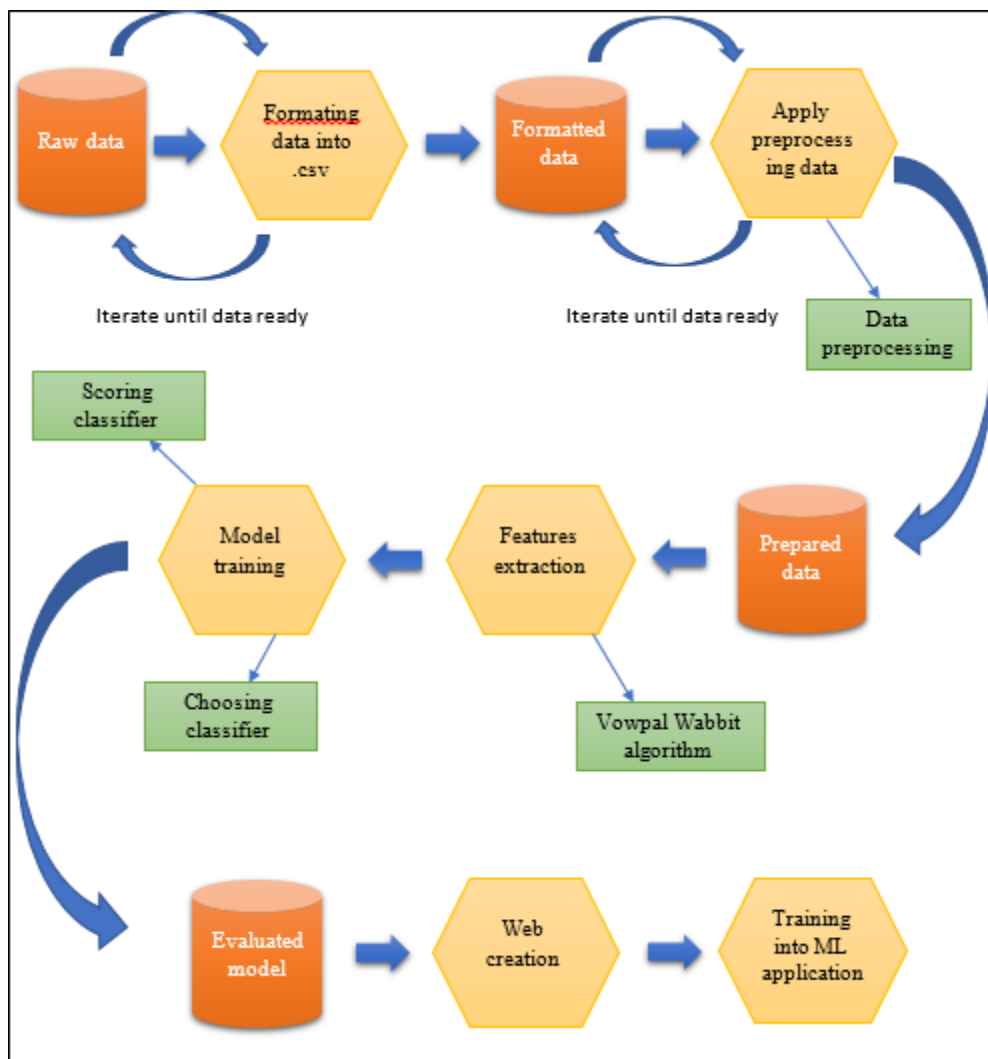


Figure 8: Data model flow

## **DESCRIPTION**

The data model flow is essential to this project to show the structured of the project on how it should be built and how the process is related to each other. It helps to make organize the process of project smoothly and clearly. Based on the framework, Azure ML Studio is used as the platform to develop the project. First, study and discovered all the functionality on Azure ML to make sure the project can achieve its objectives. After that, make sure to download and used the real existing dataset from machine learning data repository as training and testing data. Preprocessing data start with reformat the dataset into 2 separate files which is training.csv and testing.csv format. Then, upload the formatted dataset into Azure ML under dataset function/menu and drag them onto the workspace to visualize the data. Choose any desired filters to clean the raw data such as “remove numbers” filter. In features extraction, we will transform the data so that it can be used to train the classifier by using Vow pal Wabbit algorithm. First, use the feature hashing step to change the message hashing bit size. This step is important to extract all pairs of bigrams, compute an 8-bit hash for each bigram and create a new column for each hash in the output dataset. Model training step include 2 steps which is picking a classifier and scoring the classifier. Two-Class Logistic Regression is used to predict the probability of spam detection either it is spam or ham. After the data have been trained, the model needs to be tested to evaluate its accuracy and overstrain the model so that it memorizes the data. Web service is set up so that the model can be used. First, select only message column by using Select Columns in Dataset step so that the data can be tested in the web service. After the web service is up, paste any message into the form to classify the message.

## **SUMMARY**

The result that will be taken is the type of classification either spam or ham, Methodology is one of the most important roles in system and application development. There also a lots of different software development methodology that available and can be used to develop any kind of application. The right methodology can help the project to be done according to the specified time. The activities in each phase in the methodology are explained so that it can be understood easily. The detection accuracy, an



## **CHAPTER : 4**

### **TECHNOLOGY USED**

Among all the techniques developed for detecting and preventing spam, filtering email is one of the most essential and prominent approaches. Several machine learning and deep learning techniques have been used for this purpose

**1: Naïve Bayes :** Naïve Bayes is a simple learning algorithm that utilizes Bayes rule together with a strong assumption that the attributes are conditionally independent, given the class. While this independence assumption is often violated in practice, naïve Bayes nonetheless often delivers competitive classification accuracy.

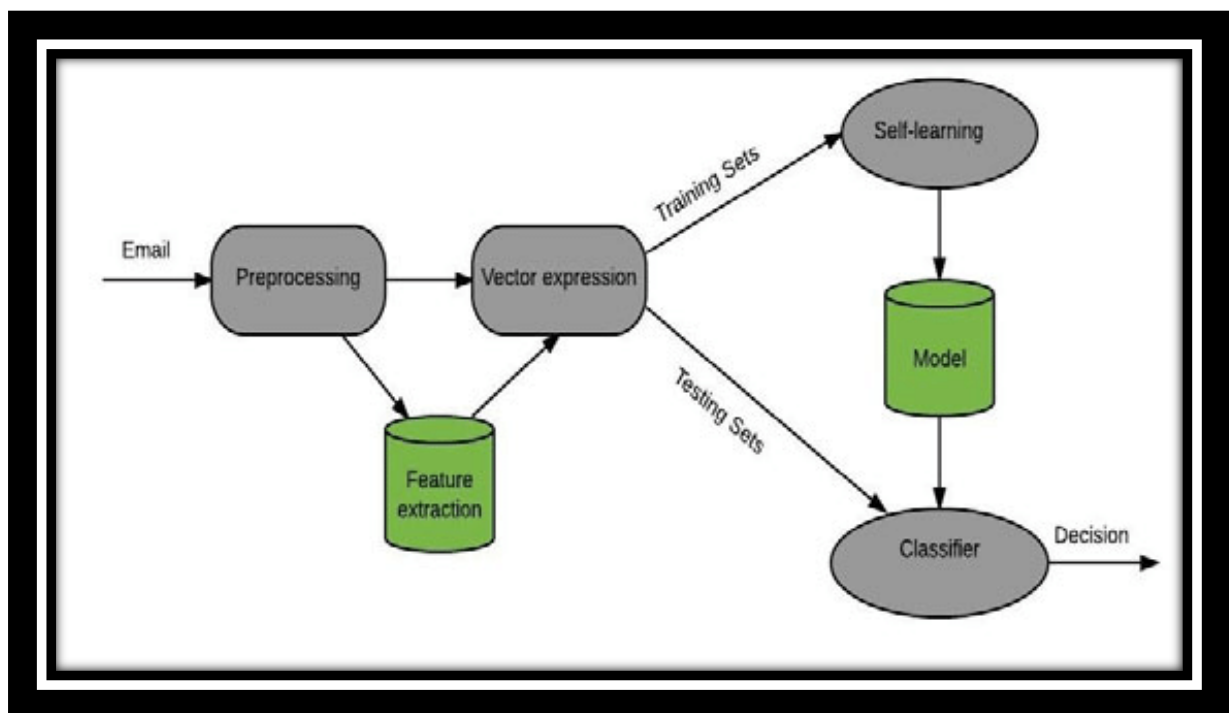
**2: Decision trees :** Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

**3: Neural networks :** A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

**4: Random forest :** Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a “forest.” It can be used for both classification and regression problems in R and Python.

## CHAPTER: 5

### DIAGRAMS



## **CHAPTER: 6**

### **CONCLUSION**

The performance of a classification technique is affected by the quality of data source. Irrelevant and redundant features of data not only increase the elapse time, but also may reduce the accuracy of detection. Each algorithm has its own advantages and disadvantages as stated in Chapter 2. As state before, supervised ML is able to separate messages and classified the correct categories efficiently. It also able to score the model and weight them successfully. For instances, Gmail's interface is using the algorithm based on machine learning program to keep their users' inbox free of spam messages. During the implementation, only text (messages) can be classified and score instead of domain name and email address. This project only focus on filtering, analyzing and classifying message and do not blocking them. Hence, the proposed methodology may be adopted to overcome the flaws of the existing spam detection.

## **REFERENCES**

- 1:- Anitha, PU & Rao, Chakunta & , T.Sireesha. (2013). A Survey On: E-mail Spam Messages and Bayesian Approach for Spam Filtering. *International Journal of Advanced Engineering and Global Technology (IJAEGT)*. 1. 124- 136.
- 2:- Attenberg, J., Weinberger, K., Dasgupta, A., Smola, A., & Zinkevich, M. (2009, July). Collaborative email-spam filtering with the hashing trick. In *Proceedings of the Sixth Conference on Email and Anti-Spam*.
- 3:- Awad, W. A., & ELseuofi, S. M. (2011). Machine learning methods for spam e-mail classification. *International Journal of Computer Science & Information Technology (IJCSIT)*, 3(1), 173-184.
- 4:- Barnes, J. (2015). Azure Machine Learning. *Microsoft Azure Essentials*. Isted, Microsoft.
- 5:- Chang, M. W., Yih, W. T., & Meek, C. (2008, August). Partitioned logistic regression for spam filtering. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 97- 105). ACM
- 6:- Çıltık, A., & Güngör, T. (2008). Time-efficient spam e-mail filtering using n- gram models. *Pattern Recognition Letters*, 29(1), 19-33.
- 7:- Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1), 23.
- 8:- Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007, July). Transferring naive bayes classifiers for text classification. In *AAAI* (Vol. 7, pp. 540-545).
- 9:- Fishkin, R. (2015, November 06). Spam Score: Moz's New Metric to Measure Penalization Risk. Retrieved from <https://moz.com/blog/spam-score-mozs-new-metric-to-measure-penalization-risk>
- 10:- Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7), 10206- 10222.