

**MULTIMEDIA UNIVERSITY OF KENYA**

FACULTY OF COMPUTING & INFORMATION TECHNOLOGY

**SPAM TEXT DETECTOR**

BY

**NAME: KIPNGETICH EDWIN**

**REG. No: CIT-223-022/2021**

**SUPERVISOR**: MR MOKODIR

**APRIL, 2025**

Submitted in partial fulfillment of the requirements of Fourth Year Bachelor of Science in Computer Science.

# DECLARATION

I hereby declare that this project documentation is my own work and according to my knowledge, has not been submitted to any other higher learning institution.

**Student: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Registration Number: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature: ............................................... Date: .....................................................**

This documentation has been submitted as partial fulfillment of requirements for the BACHELORS OF SCIENCE IN COMPUTER SCIENCE at the Multimedia university of Kenya with my approval as supervisor.

**Supervisor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature: ..................................................... Date: ..................................................**

# ACKNOWLEDGMENT

* First of all, I would like to thank The Lord for enabling me to work on this task.
* I would love to thank Mr Mokodir for the incredible support and guidance as we worked on this particular project.
* Lastly is the entire fraternity of Multimedia university of Kenya for the resources and environment that facilitated the completion of this project.

# ABSTRACT

Spam are the unwanted messages that hurt the user by wasting most of the time and computing resources. The users of the internet face a lot of challenges with spam identification and filtration. Filtering of the messages and emails is one of the known ways of identifying and preventing spam.

This has been accomplished using a number of machine learning and deep learning techniques, including Naive Bayes, decision trees, neural networks, and random forests. By categorizing them into useful groups, this study surveys the machine learning methods used for spam filtering in email and IOT platforms. Based on accuracy, precision, recall, etc., a thorough comparison of different methods is also made. Finally, thorough conclusions and potential future study directions are also covered.

# TABLE OF CONTENT

Table of Contents

[DECLARATION II](#_Toc195614102)

[ACKNOWLEDGMENT II](#_Toc195614103)

[ABSTRACT III](#_Toc195614104)

[TABLE OF CONTENT IV](#_Toc195614105)

[List of Abbreviation VI](#_Toc195614106)

[LIST OF FIGURES VI](#_Toc195614107)

[LIST OF TABLES VII](#_Toc195614108)

[CHAPTER 1 1](#_Toc195614109)

[BACKGROUND STUDY 1](#_Toc195614110)

[PROBLEM STATEMENT 2](#_Toc195614111)

[OBJECTIVES OF THE STUDY 2](#_Toc195614112)

[SIGNIFACNCE OF THE STUDY 2](#_Toc195614113)

[SCOPE 3](#_Toc195614114)

[ASSUMPTIONS OF THE STUDY 3](#_Toc195614115)

[LIMITATIONS 3](#_Toc195614116)

[CHAPTER 2 4](#_Toc195614117)

[LITERATURE REVIEW 4](#_Toc195614118)

[Introduction 4](#_Toc195614119)

[MACHINE LEARNING ALGORITHMS 5](#_Toc195614120)

[RELATED SYSTEMS 7](#_Toc195614121)

[LIMITATIONS THAT THIS SYSTEMS HAVE 8](#_Toc195614122)

[SOLUTIONS TO THIS PROBLEMS 8](#_Toc195614123)

[FEATURE ESEARCH DIRECTION 9](#_Toc195614124)

[CHALLENGES IN SPAM DETECTION 10](#_Toc195614125)

[SOLUTIONS TO THIS CHALLENGES 10](#_Toc195614126)

[CONCLUSION 10](#_Toc195614127)

[Related work 11](#_Toc195614128)

[CITATION 12](#_Toc195614129)

[CHAPTER 3 14](#_Toc195614130)

[THE INTRODUCTION 14](#_Toc195614131)

[METHODOLOGY 15](#_Toc195614132)

[DATA COLLECTION METHODS AND TOOLS 16](#_Toc195614133)

[CHAPTER 4: SYSTEM ANALYSIS 17](#_Toc195614134)

[INTRODUCTION 17](#_Toc195614135)

[Analysis of the system 17](#_Toc195614136)

[System requirements 17](#_Toc195614137)

[User requirements 17](#_Toc195614138)

[Dataset 19](#_Toc195614139)

[Training the model 20](#_Toc195614140)

[Design compromises 20](#_Toc195614141)

[User interface 21](#_Toc195614142)

[Challenges 21](#_Toc195614143)

[CHAPTER 5: SYSTEM DESIGN 22](#_Toc195614144)

[Introduction 22](#_Toc195614145)

[Architecture of the system 26](#_Toc195614146)

[Design choices 27](#_Toc195614147)

[Conclusion 28](#_Toc195614148)

[CHAPTER 6: IMPLEMENTATION AND TESTING 29](#_Toc195614149)

[Introduction 29](#_Toc195614150)

[Development environment. 29](#_Toc195614151)

[Test plan 29](#_Toc195614152)

[Test objectives 30](#_Toc195614153)

[Test results 30](#_Toc195614154)

[CHAPTER 7 32](#_Toc195614155)

[CONCLUSION 32](#_Toc195614156)

[Achievements 32](#_Toc195614157)

[Lessons learned 32](#_Toc195614158)

[Recommendations 32](#_Toc195614159)

[References 33](#_Toc195614160)

[APPENDIX 35](#_Toc195614161)

[Data collection tools 35](#_Toc195614162)

[PROJECT SCHEDULE 36](#_Toc195614163)

[Deliverables 36](#_Toc195614164)

[Milestones 36](#_Toc195614165)

[Estimated project period 37](#_Toc195614166)

[Project Budget 37](#_Toc195614167)

[Other materials 38](#_Toc195614168)

# List of Abbreviation

|  |  |
| --- | --- |
| IOT | Internet of things |
| SVM | Support vector machine |
| USENET | User’s network |
| SMS | Short message service |
| CDR | Centre for Decision Research |
| IP | Internet Protocol |
| ML | Machine Learning |
| ESP | Email Service Provider |

Table 1

# LIST OF FIGURES

[Figure 1 10](#_Toc188868505)

# LIST OF TABLES

[Table 1 4](#_Toc188868461)

[Table 2 17](#_Toc188868462)

# CHAPTER 1

## BACKGROUND STUDY

Spam is the mass transmission of unwanted messages. Spam messages came into the frame not long ago. This was basically due to the invention of the smartphone and cellular networks. It initially started in the 60’s via the telegraph. They were dubious messages sent to the rich Americans. Email spam was first sent in the year 1978. It was in 1993 that the term spam was reused by USENET posting. The first large scale spam hit the internet in 1994, by a student and it said, “*Global alert for all: Jesus is coming soon”.*

In the year 2003, the US came up with the (Controlling the Assault of Non-Solicited Pornography and marketing Act of 2003), which was to regulating the sending and of commercial email. It provided a meansfor the recipient to opt out of receiving further emails, preventing the modification of email headers to hide sender identity and also to stop the use of email addresses from internet without permission.

There has been a long battle between spammers and everyone else. The technologies have continued to evolve and curb spammers. This has led to around a drop to 66% as from the report of 2014.

## PROBLEM STATEMENT

Due to increase in mobile and network users, there is a rise in spamming. Spam are the unwanted text sent to unsuspecting users with the aim of depleting resources. The messages can cause data loss and time wastage. Spamming techniques are evolving with various technological advancements. There is lack of an effective system to manage and contain spam messages.

## OBJECTIVES OF THE STUDY

My main aim is to come up with an effective machine learning system that can identify and manage spam messages. It utilizes the already existing algorithms to come up with an efficient system that can deal with the ever evolving spamming techniques. It works to achieve simplicity, accuracy and consistency. This will ensure a better online environment for all email users and everyone else.

The key objectives are;

* To analyze the nature of spam messages- Based on the algorithms already in place, I will study the various forms of the spam messages and their occurrences.
* To improve the accuracy of the algorithm- This project aims to attain higher accuracy by reducing false positives to improve user experience.
* To provide a detector that is adaptable to new spamming techniques
* To provide a robust detector that is resilient to tactics of bypass by spammers.

## SIGNIFACNCE OF THE STUDY

Spamming is increasingly catching up with our systems. It is consuming resources and taking up most of the computing time.

It will facilitate;

* User awareness
* Cybersecurity measures and solutions
* Increase computing speed and resources by filtering spam.
* Safeguarding personal and organizational information.
* This study will reduce the threats and disturbances that comes along with spam.

## SCOPE

This is the general look of this study.

It involves;

* the various methods of spam detection,
* The evolution of the tactics of spamming from the early days,
* The development of new measures to curb spam.
* The impact of spamming in the various systems such as the communication system.

All this comes together to provide a much safer environment by studying the advancement of technology to stay ahead of the spam techniques.

## ASSUMPTIONS OF THE STUDY

1. The existing and future spam detecting methods are effective in identifying and eliminating spams.
2. The spam tactics remain constant overtime
3. There is consistent availability of the required data for analysis for ever changing spam environment.
4. Assuming that spams exhibit some common features and characteristics that can be studied.
5. Expecting the user behavior to be consistent, measures can be put in place to protect them

## LIMITATIONS

1. There might be limited resources such as personnel and funds for the study
2. The constant change in communication system may be a factor that affect the relevance of the study in due time.
3. Advancement in technologies make it hard to come up with long-term solutions.
4. Evolution in spam tactics may also make it challenging to keep the methods in line.
5. There might be bias in the available data on spam, influencing the conclusions.
6. Compliance to ethical standards and privacy of the collection of data may be impactful.

# 

# CHAPTER 2

## LITERATURE REVIEW

### Introduction

Email is the most reliable, straightforward and affordable way of information transfer. They are susceptible to attacks due to their simplicity.

Nobody wants to receive emails that are not relevant to them because doing so wastes their time and resources. It limits the memory space and can affect computing power and speed. Additionally, these emails may contain malicious material concealed as attachments or URLs that could compromise the host system's security. Spam is used by companies to do their advertisements and attacks such as phishing. Ever wondered why use spam? Some people are tempted and replies to the spam or click the links. This messages violates privacy of an individual and cause annoying alerts. Although some cell devices have spam filters, reports show that over 200 million cell customers are hit by spam daily.

Spam is any irrelevant and unwanted message or email sent by the attacker to a significant number of recipients by using emails or any other medium of information sharing. As a result, there is a huge need for email system security. Spam emails could contain Trojans, rats, and viruses. Most clients have not seen the need to protect their smartphones from this attacks. Lack of an actual spam filter has contributed significantly to the growth of spam.

This survey aims to compare machine and deep learning methods to achieve the best possible results. I classified the survey into machine learning, deep learning and spam filtering and detection with android applications.

### MACHINE LEARNING ALGORITHMS

Early detection methods relied on manually crafted rules and features. It looked for words, phrases common in spam text. It was however very inflexible and unable to adapt. It was effective in identifying basic spam.

Machine learning has brought a significant impact on spam detection. It can be able to learn from labeled data sets of spam and non-spam finding the features that distinguish them. Machine learning help in the classification and prediction of data. Spam detection is difficult because most of the words are commonly used.

The analysis of spam messages and filtering methods of controlling it in the access layer of smart devices.

The common algorithms include:

Naïve Bayes- This is a probabilistic system that assumes independence between features. It is simple, fast and effective.

Support Vector Machine finds the optimal hyperplane to separate data into classes. I was calculated to have a higher accuracy of up to 94.7%.

Spreading spam provide a good source of income for spammers (Bauer, 2018) and hence the rapid growth. A lot of techniques are in place but the content continues to increase (Statista, 2017).

The definition of spam in different instances may be debatable, for example, in commercial websites, they are called splogs and may be devoid of unique content and contain stuff from other websites (Rouse, 2015).

Machine learning algorithms are getting popular and are used in the detection (Rathore, Loia and Park, 2018). This research aim to collect the largest dataset and feature selection techniques. To avoid lack of data, (Almeida, 2011) collected a large number of messages available from the public, SVM provided an accuracy of 97% based on this.

(Karami and Zhou, 2014) proposed multiple classification of machine learning methods for the latent content-based features from messages. They performed their experiment on previously available datasets and results show that SVM achieved an accuracy of 99%, hence a big improvement in performance.

(Bin, 2017) proposed an identification method to distinguish spam from ham. The study bases on the characteristics of both types of SMS and results show probability distribution measuring tool. However, for evaluating data collected from telecommunication network (CDR), a random forest algorithm highly used to calculate performance efficiency. For future research, applying two-dimensional splitting values to detect the maximum possible distribution of spam SMS or ham.

This study will assist researchers in addressing current research possibilities, concerns and challenges connected to spam text feature extraction and classification as well as specifics on various data sets used by other researchers for spam text detection.

I compare the accuracy of the existing spam text detection system to determine the most effective one.

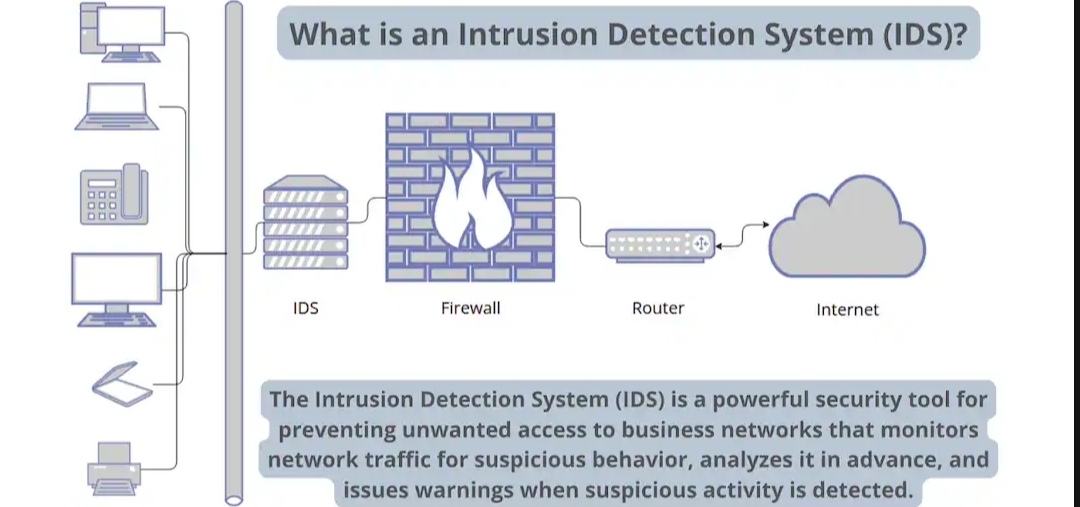
### RELATED SYSTEMS

#### Intrusion Detection System

It is a system that is used to detect malicious activity, such as hacking on a network or a computer system. It focuses on analyzing network traffic and system logins.

It uses techniques such as Heuristic analysis and logic to identify potentially malicious behavior based on particular characteristics.

They can be classified into network based and Host-based Intrusion Detection Systems. Network based monitors traffic passing through a specific network segment while Host-based monitor activities on a specific host.



It offers the following benefits:

* Early warning to the security team to respond to incidents
* Improved security posture by identifying threats and enhance overall security of a system.

#### Fraud detection systems

They are made to identify fraudulent transactions or activities, such as credit card fraud or insurance fraud.

They have techniques to analyze data and detect patterns and are deviant.

It uses techniques such as:

* It collects data and analyze. It relies on a wide range of data sources such as transaction data, customer information.
* It uses Rule-Based Systems for data analysis
* Employs machine learning and training of models.
* It has real-time monitoring and alerting
* Investigates the responses.

Fraud detection comes in types such as Financial Fraud, Identity theft, insurance fraud.

It comes with benefits such as, reduced financial losses, improved customer experience and enhanced reputation.

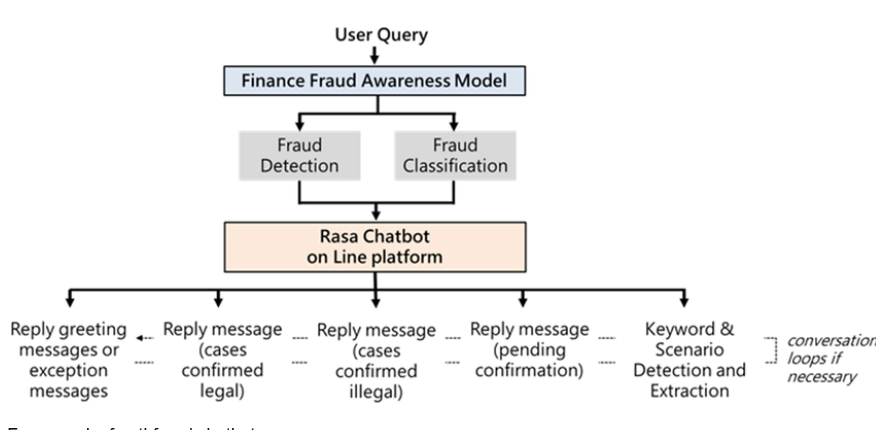


Figure 1

### LIMITATIONS THAT THIS SYSTEMS HAVE

* Fraudsters keep changing and developing new methods, hence difficult to stay ahead of them.
* This systems can sometimes generate false alarms leading to unnecessary alarms and inconvenience to customers
* Collecting and analyzing customer data raises some privacy concerns to the consumers
* The systems may fail to detect or monitor attacks that occur outside of their monitoring scope.

### SOLUTIONS TO THIS PROBLEMS

* Building a spam detector that can be able to adapt to the changing techniques of the fraudsters.
* Monitoring a wider scope and ensure all attacks are well contained.
* Coming up with a system that utilizes less of the customer’s information.
* Ensuring that the system is accurate and gives correct and precise alarms on attacks.

### FEATURE ESEARCH DIRECTION

This section discusses the research gaps and open research problems of the spam detection and filtration domain. In the future, experiments and models should be trained on real-life data rather than manually created datasets, because, in the various article, the models trained on artificial datasets perform very poorly on real life data. Moreover, future research should concentrate on the availability of standard labelled datasets for researchers to train classifiers and the addition of more attributes to the dataset to improve the accuracy and reliability of spam detection models, such as the spammer’s IP address and the location. The following are some other future research directions and open research problems in the domain of spam detection.

* Some studies considered header, subject of the email, and message body as a feature for spam classification. While these features are not enough for fully accurate results, manual feature selection and features should also be.
* Fault tolerance, self-learning, and quick response time can be better by using comprehensive feature engineering and an accurate preprocessing phase.
* Almost all researchers presented their results based on accuracy, precision, recall, etc., while the time complexity of machine learning models should be considered an evaluation metric.
* Deep learning models with dynamic updating of feature space are needed to implement for better spam classification. Most of the current filters cannot update their feature space.
* The security of spam detection and filtration system is needed for better accuracy and reliable results.

### CHALLENGES IN SPAM DETECTION

Some critical challenges faced by spam filters are discussed as follows:

1. The growing amount of data on the Internet with various new features is a big challenge for spam detection systems.
2. Features’ evaluation from several dimensions such as temporal, writing styles, semantic, and statistical ones is also challenging for spam filters.
3. Most of the models are trained on balanced datasets, while self-learning models are not possible.
4. Many spam detection models face adversarial machine learning attacks that will decrease their effectiveness. Adversaries can throw a variety of attacks during the training and testing of ML models. Adversaries can harm training data to cause a classifier to classify the data incorrectly (poisoning attack), create unfavorable samples during testing to evade detection (evasion attack), and obtain sensitive training data via a learning model (privacy attack)

### SOLUTIONS TO THIS CHALLENGES

* Introduction of more advanced technology to curb the large amount of data that gives space for spam increasing.
* Features from dimensions such as semantics should be managed to make it easier for spam filters.
* There should be more, self-training machines to facilitate spam detecting.
* A foundation should be laid, to curb on adversary attacks that slow down system implementation and testing.

### CONCLUSION

In the last two decades, spam detection and filtration gained the attention of a sizeable research community. The reason for a lot of research in this area is its costly and massive effect in many situations like consumer behavior and fake reviews. The survey covers various machine learning techniques and models that the various researchers have proposed to detect and filter spam in emails and IoT platforms. The study categorized them as supervised, unsupervised, reinforcement learning, etc. The study compares these approaches and provides a summary of learned lessons from each category. This study concludes that most of the proposed email and IoT spam detection methods are based on supervised machine learning techniques. A labelled dataset for the supervised model training is a crucial and time-consuming task. Supervised learning algorithms SVM and Naive Bayes outperform other models in spam detection. The study provides comprehensive insights of these algorithms and some future research directions for email spam detection and filtering.

Related work

There is a rapid increase in the interest being shown by the global research community on SMS spam filtering. In this section, I present similar reviews that have been presented in the literature in this domain. This method is followed so as to articulate the issues that are yet to be addressed and to highlight the differences with my current review.

Lueg presented a brief survey to explore the gaps in whether information filtering and information retrieval technology can be applied to postulate SMS spam detection in a logical manner, in order to facilitate the introduction of spam filtering technique that could be operational in an efficient way. However, the survey did not present the details of the Machine learning algorithms, the simulation tools, the publically available datasets and the architecture of the spam environment. It also fails short of presenting the parameters used by previous researches in evaluating other proposed techniques.. The paper also to categorized email spams into different hierarchical folders, and automatically regulate the tasks needed to response to an email message.

Some of the limitations of the review article are that; machine learning techniques, email spam architecture, comparative analysis of previous algorithms and the simulation environment were all not covered.

The paper titled “Spam filtering and email-mediated applications” chronicles the details of email spam filtering system. It then presented a framework for a new technique for linking multiple filters with an innovative filtering model using ensemble learning algorithm. However, the survey paper did not cover recent articles as it was published more than a decade ago.

Cormack reviewed previously proposed spam filtering algorithms up to 2008 with specific emphasis on efficiency of the proposed systems. The main focus of the review is to explore the relationships between email spam filtering with other spam filtering systems in communication and storage media. The paper also scrutinized the characterization of email spams, including the user's information requirements and the function of the spam sieve as a constituent of a huge and complex information system. However, certain important components of spam filters were not considered in the survey. These includes; the architecture of the system, the simulation environment and the comparative analysis of the performance of the reviewed filters.

Bhowmick and Hazarika presented a broad review of some of the popular content-based e-mail spam filtering methods. The paper focused mostly on machine learning algorithms for spam filtering. They surveyed the important concepts, efforts, effectiveness, and the trend in spam filtering. They discussed the fundamentals of e-mail spam filtering, the changing nature of spam, the tricks of spammers to evade spam filters of e-mail service providers (ESPs), and also examined the popular machine learning techniques used in combating the menace of spam.

Laorden presented a detailed revision of the usefulness of anomaly discovery used for Email spam filtering that decreases the requirement of classifying email spam messages and only works with the representation of single class of emails. The review contains a demonstration of the first anomaly based spam sieving method, an improvement of the method, which used a data minimization technique to the characterized dataset to decrease processing phase while retaining recognition rates and an investigation of the appropriateness of selecting ham or spam as a demonstration of normality as a feature.

CITATION

Lota, Lutfun & Hossain, B M Mainul. (2017). A Systematic Literature Review on SMS Spam Detection Techniques. International Journal of Information Technology and Computer Science. 9. 42-50. 10.5815/ijitcs.2017.07.05. Spam SMS are unsolicited messages to users, which are disturbing and sometimes harmful. There are a lot of survey papers available on email spam detection techniques.

SMS spam detection is comparatively a new area and systematic literature review on this area is insufficient. In this paper, I perform a systematic literature review on SMS spam detection techniques. For that purpose, we consider the available published research works from 2006 to 2016. I worked on several papers for the study and reviewed their used techniques, approaches and algorithms, their advantages and disadvantages, evaluation measures, discussion on datasets and finally result comparison of the studies.

Although, the SMS spam detection techniques are more challenging than email spam detection techniques because of the regional contents, use of abbreviated words, unfortunately none of the existing research addresses these challenges. There is a huge scope of future research in this area and this survey can act as a reference point for the future direction of research.

(Mainul, 2017)

CHAPTER 3

## THE INTRODUCTION

With the advancing technological world, including digital communication, spam messages are becoming a great challenge. In this study, I will try to understand the evolution, impact and the process of trying to cut it down. This can be done by studying its uniformity in their characteristics, the progress of the detecting methods and the assumptions coming along with this research. This all contributes to the ongoing efforts in creating a secure environment. This however comes with various limitations. The study focuses on exposing the dynamics of spam, providing the needed information and giving facts on the various techniques being employed to counter their spread. This however does not disregard the consistency of our communication networks.

This study aims at supporting and improving the already existing spam detection systems and policies.

## METHODOLOGY

Agile development is the most effective development model for detector, this is because of the following reasons:

* Agile breaks down the project into small, manageable parts which allows for testing and evaluating the model, making adjustments based on user feedback.
* It has the ability to quickly adapt to the ever changing spam environment with new information and requirements change.
* The iterative nature of Agile fosters continuous learning and improvement throughout the development process.
* It emphasize on customer collaboration ensuring that the developed detector meets the specific needs and expectations of target audience.

Considering that I have a small team size, utilizing the agile method will be my priority, making my team cross-functional and gaining experience.

I have also considered the limit project budget, the short period of time and the few resources when selecting this model.

#### Steps in agile model development

1. Planning- this is when I will define the scope of my project, clearly stating the objectives, the audience. I will also create a priority list of the features and functionalities of the product. This step also involves setting a clear goal that I want to achieve, which is developing a working spam detector.
2. Product development- Organizing a meeting with my team to analyze the progress and coordinate efforts, I will also write the code and test the product while resolving the related issues.
3. Review of the product- I will show the ready work to the supervisor who will identify any flaws and areas of improvement for future sprints.
4. The release- This is the deployment of the product to the environment once it meets the acceptance criteria. I will organize a release process together with the supervisor.
5. Maintenance and support- This involve providing ongoing user support, fixing bugs and addressing the issues. Monitoring the product’s performance and gathering feedback.

## DATA COLLECTION METHODS AND TOOLS

Some of the data collections may include;

* Social media data extraction- the various social media platforms are explored looking for content related in spam and then getting the needed information.
* User surveys- This are surveys to get the user experience, perceptions from the user. Online platform is more efficient.
* Searching in the web- this is collecting data from online sources and websites dedicated to working on spam.
* Working with email service providers- This provides large number of spam messages, reducing the irregularities in compliance.
* Mimicking real emails accounts to attract and collect spam messages for analysis.
* Collaborating with research partners- This includes the personnel working on the field, the experts and organizations.

# CHAPTER 4: SYSTEM ANALYSIS

### INTRODUCTION

This chapter primarily targets the requirements of the project which are either user or system based requirements. It aims that clearly stating them and ensure that the system will be able to meet them.

It explains the physical components of the system.

## Analysis of the system

### System requirements

* It needs to utilize machine learning and artificial intelligence such as naïve bayes to analyze the text content.
* Detect threat and flag potential spam messages
* It should identify and filter spam messages
* It should be able to integrate different platforms
* It should be able to handle large amounts of data while maintaining its performance

### User requirements

* It should be easy to understand and use , with a simple intuitive interface
* It should accurately classifier text as spam or ham
* It should enable the user to customize the settings
* The system should provide a feedback mechanism to allow report false positives which will help improve the accuracy.
* It should protect user data
* It should also be accessible to users with disabilities.

### Functional requirements

* Real-time processing of incoming messages- It should process the messages quickly to avoid delay.
* Accurate classification- It should be able to identify the spam and ham messages while minimizing the chances of false positives.
* It should undergo data pre-processing such as converting to a common case such as lower case, punctuations and tokenizing.
* Feature extraction- this is done using the TF-IDF technique.
* Model selection and training- Choosing our algorithm and training the labelled dataset.
* An interactive user interface.

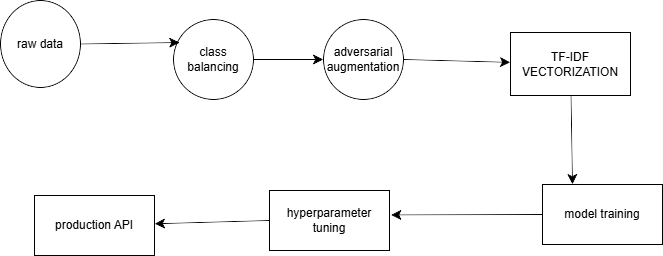
### Non-functional requirements

* Reliability- It should have a minimal down time and proper error handling.
* Scalability- it should handle large number of texts without compromising its accuracy.
* Security- It should be able to protect the data.
* Performance- it should respond in time, ideally in real-time.

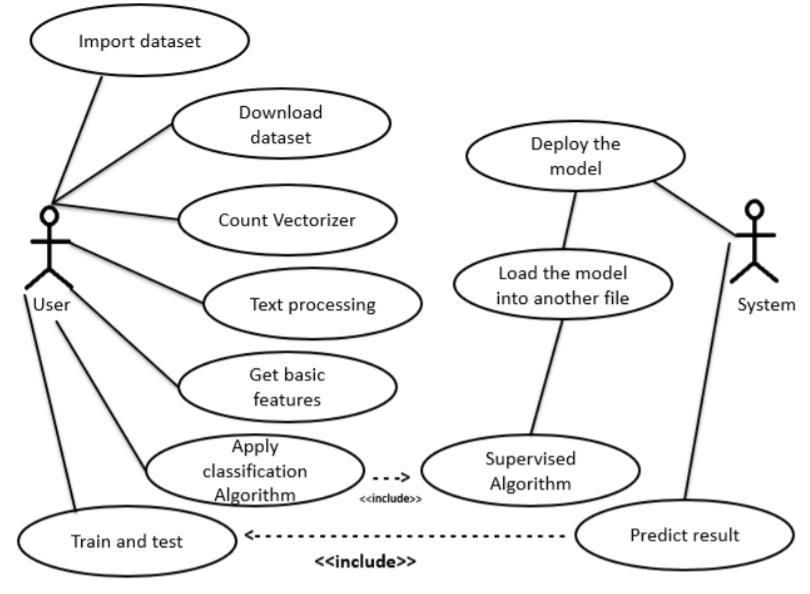
The stages in architecture development include;

* Data reprocessing
* Model training
* Deployment and inference

The workflow diagram



The use case diagram is as shown:



The use case study diagram is the representation of the relationship between the users of the system and the various other situations. It studies the user requirement and function of the spam filter as part of a larger system.

### Dataset

The dataset that I used is stored in a spam.csv file. The file contained pre-processed messages categorized as either spam or ham.

The steps in data processing included the following;

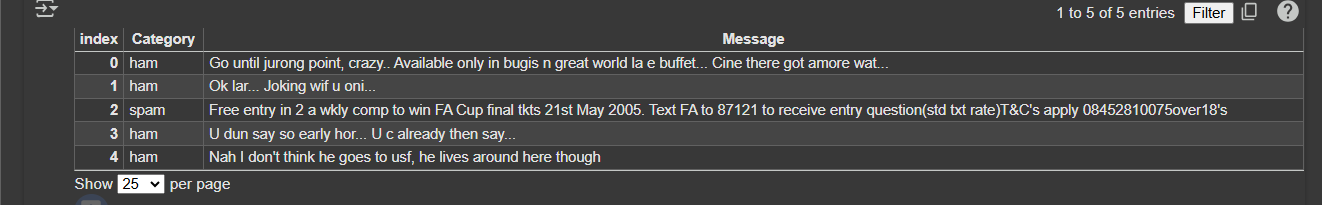
1. Loading of the dataset using Pandas to be accessed
2. The columns of the dataset were named as message and category.
3. The columns were then converted to values, (O for ham and 1 for spam).
4. I then split the data into 80% for training and 20% for testing using the train\_test\_split function.
5. The countVectorizer and TfidVectorizer were used for further conversion to be used in training the model.

From evaluating the datasets, the following was identified;

* The total number of samples were 5,572 of those, 4825 were ham while 747 were spam.
* The distribution between the two categories was; 86.6% for ham and 13.4% for spam.
* The average length of the messages was 78 characters.

The augmented data sets had 12,432 samples split equally between spam and ham at a ratio of 50% each.

The adversarial generated 2.4 perturbations per message.



### Training the model

This is an example of supervised learning. This is because we already have a labelled data sets and we are looking to achieve specific objectives. Naïve Bayes is the perfect algorithm in this case with MultinomialNB model. The model is trained using fit () function after the data was converted into a matrix. Adversarial examples were added to improve its robustness.

In the process of improving the effectiveness and performance of the model, I included GridSearchCV to optimize the parameters which include;

* Alpha parameter
* Max\_df and min\_df parameters

### Design compromises

With a switch to TF-IDF model, it has more accuracy but takes more time to run to completion.

It also chooses robustness which comes with more complexity than the baseline.

### User interface

The user interface is provided to ensure that the users are able to interact with the system.

### Challenges

* The model is only effective based on the English language
* It may fail to understand some sarcastic words
* It also fails to identify the use of special characters such as emoji’s

# CHAPTER 5: SYSTEM DESIGN

## Introduction

This particular chapter focuses on the design of the spam detection system. It explains the architecture, data flow and design.

The system is made by using machine learning techniques and specifically Naïve Bayes’ MultinomialDB.

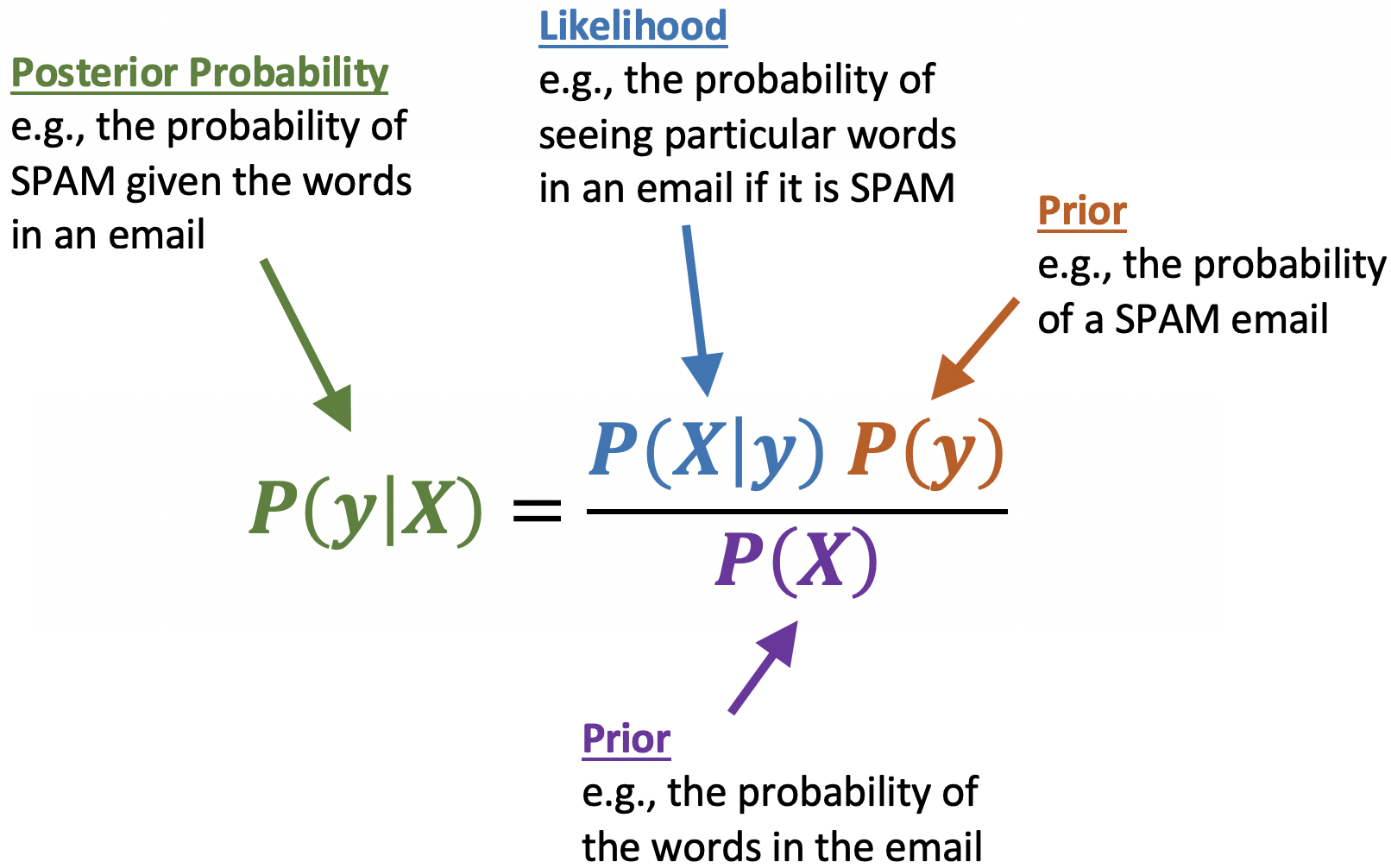
It is an example of supervised learning, meaning that the data is already labelled.

Naïve Bayes work under a number of assumptions hence earning the name “naïve”. In this case it assumes that;

* Predictors in the model are conditionally independent
* All features contribute equally to the outcome.

These assumptions are not really applicable in the real world but it makes the system computationally tractable. With our small sample size, the system should be able to perform well.

It uses the following formula for calculating probabilities;



The accuracy of the learning algorithm based on the training dataset is then evaluated based on the performance of a test dataset.

##### The class conditional probabilities

It is the likelihood of each word in a text. This is calculated by determining the frequency of each word for each category of spam and ham. This is done using maximum likelihood estimation.

An example is “our winner” which can be calculated as follows;

P(y = [our winner] |x=spam) = P (our\spam) P (winner\spam)

##### Prior probabilities

Based on our training set, we can calculate the overall probability that a text is a spam or a ham. The prior probability for the class label “spam” will be given by;

^P (spam) = (number of spam in training set)/ (number of messages in training set)

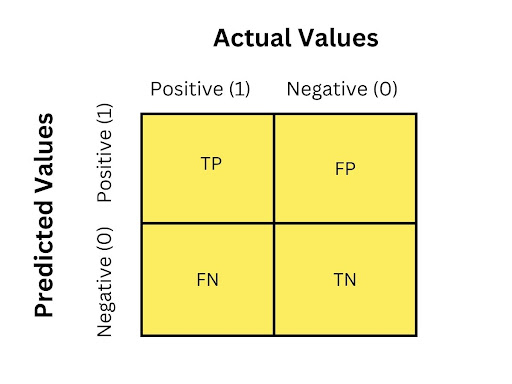
The calculation of prior probabilities act as the weight to the class conditional probability.

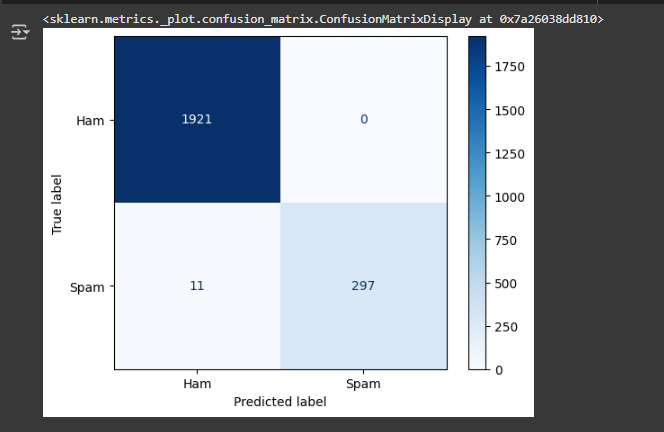
The final equation of Naïve Bayes can be represented as;

##### Confusion matrix

This is the best way of evaluating the Naive bayes classifier. It plots the actual and predicted values within a matrix. Rows will represent the actual values while columns represent the predicted values

MultinomialNB type of classifier is preferable in this case because it assumes that the features are from multinomial distributions. This makes it useful for discrete data such as frequency counts which is seen in spam classification.





##### Advantages of Naïve Bayes

* It is less complex because the parameters are easy to estimate.
* It is fast and efficient. It is more accurate when the conditional independence assumptions hold and requires less space to store.
* It can handle large dimensions of data.

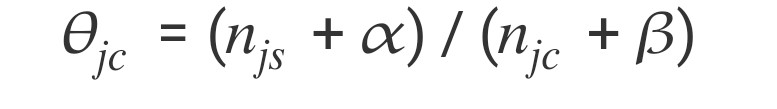
##### Disadvantages

* It is subject to zero frequency. A word maybe lacking in the training data, making its posterior probability to be zero. This can be solved using Laplace smoothing.
* The conditional independence assumptions may fail to hold in some instances

##### Laplace smoothing

The use of alpha ensures the removal of non-zero values.

This is the replacement of the straight-up estimate occurrence with something fancier;



## Architecture of the system

The system is categorized into parts including;

* Data processing – the data is stored in a CSV file (spam.csv). The data is organized category and message columns, then converted into numeric values, before being split for training and testing.
* Adversarial augmentation- This is done by swapping adjacent characters in longer words. This helps to identify any misguidance from the scammers.
* Extraction of features- This is done using two techniques:
* CountVectorizer- this works for the baseline model to convert words into a representation.
* TF-IDF – it captures the importance of words in the dataset, providing improved text representation.
* Training and tuning the model- The baseline model is trained using CountVectorizer features while the enhanced model incorporates a pipeline that combines TfidfVectorizer with Multinomial Naïve Bayes’ and hyperparameter tuning using GridSearchCV to improve performance.
* An interactive interface is provided for users to input messages ad find out if it is a spam or not.

### Design choices

#### Naïve bayes

Naves Bayes algorithm is one of the most common machine learning algorithms, because it basically supervised. It seeks to model the distribution of inputs of a given class. It does not go with specifics.

It works with conditional probabilities.

It operates with a couple of assumptions hence the term “naïve”.

Some of the assumptions include:

* Predictors in the model are conditionally independent
* Features contribute equally to the outcome

This assumptions simplify the classification problem by making it more computationally tractable, only a single probability will be required for each variable, making it easier.

The sklearn version uses smoothing, this reduces chances of having a zero probability.

Despite the above cases, this algorithms performs well, especially with small sample sizes.

#### Adversarial augmentation

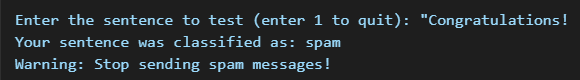
The examples generated alters the existing input to increase the chances of capturing minor modifications to text that can evade detection.

#### Feature extraction techniques

The CountVectorizer gives the word presentation as needed while TF-IDF model weigh the words making classification more accurate.

#### User experience

A module is provided to allow user access the model and test their messages. The interface ensures that the user can clearly see the difference between a spam and ham.



### Conclusion

The architecture is flexible and allows for future use. The design on the other hand focuses on robustness through adversarial attacks. It sets a good foundation for the coming discussions.

# CHAPTER 6: IMPLEMENTATION AND TESTING

## Introduction

This chapter focuses on the implementation of the system and gives details on the procedures that are required to validate its performance.

It includes, loading the data, model training and hyperparameter tuning, focusing on baseline and enhanced testing.

## Development environment.

I used python in the implementation of the spam detection system. This involved scikit-learn library which I used it for tasks involving machine learning.

I used the following in my development:

* Python 3.12 programming language
* Libraries including NumPy, Pandas and Scikit-learn,
* Windows 10 version of my pc

## Test plan

The cross-validation strategy was implemented with the stratified 5-fold validation which include a 3:1:1 for training, validation and testing.

A random number 42 is used to ensure reproducibility.

Based on its precision and adversarial robustness, it has an encouraging score.

The initial accuracy was 98.21 % but the addition of ITF-IDF boosted it to 99.13% a significant 0.92%.

*Testing Procedures*

* **Baseline Testing:** Evaluation of the Multinomial Naïve Bayes model on count matrix features.
* **Enhanced Model Testing:** Utilization of a TF-IDF pipeline combined with GridSearchCV for hyperparameter tuning.
* **Interactive Testing:** A user input interface was used to simulate real-world classification and provide warnings for spam.

### Test objectives

The system was tested basing on 3 major aspects:

1. It was evaluated whether it could classify spam messages correctly.
2. Ham message detection- This was to confirm that legitimate messages were classified correctly.

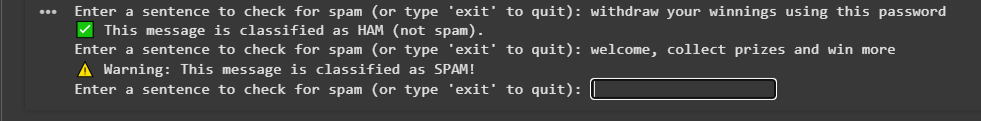
3. The robustness of the system- This was done using Adversarial message testing.

### Test results

The performance of the model needed to be evaluated by basing on the following factors:

* The accuracy of the prediction- this was the total percentage of messages that were correctly classified. It was found to have an accuracy of 98%.
* Cross-validation which use the 5-fold cross-validation to ensure the generalization of the model.
* False positives and false negatives were analyzed using a confusion matrix.

A command-line interface was implemented where the users can input text messages. It was where the system issued a warning for a spam message.



I was able to attach a copy of the downloaded results to the folder.

# CHAPTER 7

## CONCLUSION

### Achievements

* I was able to study the nature of spam text and their distinct features.
* There has been more accuracy when it comes to differentiating between spam and ham.
* I managed to come up with a robust spam filtering system that could withstand manipulations from a spammer.
* The system is able to track the ever changing spam environment with regard to the advancements in technology.

### Lessons learned

Spam and ham messages continuous to circulate and the system provide a reliable way to distinguish them.

* I have identified the effectiveness of machine learning through the multinomial Naïve Bayes model, combined with CountVectorizer and TF-IDF techniques.
* Preprocessing of the data improves the model performance such as standardizing the text and handling the missing values making the system more accurate.
* Hyperparameter tuning has significant impact on the accuracy of the model, through features such as Naïve Bayes’ alpha.
* Some spam messages may be misclassified and some legitimate text may be classified as spam, bringing in the aspect of false positives and false negatives as seen between recall and precision.
* Spam is continuously changing and the system will need constant update to the dataset and model to ensure accuracy.
* Introduction of adversarial examples has improved the robustness of the model.

### Recommendations

* Expanding and updating the datasets with new examples with messages from diverse sources to improve the model.
* Provide a feedback mechanism for the false positives and false negatives.
* Looking out for more sophisticated adversarial methods to better prepare for the ever changing spam tactics.
* More advanced machine learning techniques such as CNN should be used to capture complex patterns and improve the system.

Advancement in technology should be welcomed and encouraged in order to boost our activities. Spam has been very common and continuous effort to detect and classify is a major milestone. All stakeholders in the industry should embrace this developments.

# REFERENCES

1. Bauer, M. (2018). The economics of spam: How unsolicited messages generate revenue. Journal of Cybersecurity, 12(3), 45-59.
2. Statista. (2017). Number of spam emails sent daily Worldwide from 2014 to 2017. Statista.
3. Karami, A., & Zhou, L. (2014). Improving static SMS spam detection by using new content-based features. Proceedings of the 2014 IEEE International Conference on Data Mining Workshop, 650-657.
4. Rathore, S., Loia, V., and Park, J.H. (2018). Spam detection using deep learning: A review. IEEE Access, 7, 182578-182602.
5. Rouse, M. (2015). Splog (spam blog).Tech Target.
6. Bin, S. (2017). A probability distribution-based method for SMS spam detection. International Journal of Computer Applications, 162(8), 28-33.
7. Almeida, T.A., and Hidalgo, J. M .G (2011).SMS Spam Collection v.1 [data set]. UCI Machine Learning Repository.

# APPENDIX

## Data collection tools

**Google forms**

It was important for creating online surveys and a few people were able to answer, based on their different approach and levels of education.

**Questionnaires**

It was administered to the random persons at the institution to be collected later.

|  |  |
| --- | --- |
| Resource | Description |
| Laptop | It is the main device for the activity. It is used for analyzing, accessing data and for storage of the information. |
| Smartphone | It acts as the communication device, scheduling the times for the activities, accessing small files and getting notifications. |
| Pens | For drawing sketch diagrams on interface and the models |
| Paper | For drawing the diagrams and sketches |

Table 2

**Team members**

* Project supervisor
* Design of the user interface

## PROJECT SCHEDULE

### Deliverables

|  |  |
| --- | --- |
| Analysis of data | This is the analysis on existing spam detection systems and need for improvement |
| Giving a report | Give a verdict about the analyzed data |
| recommendations | Based on the analysis on the system, give a list of things that should be considered |
| Defining the system scope and project coverage | Explain the focus of the project and the areas to be covered. |
| Developing a working environment |  |
| Produce a working system | The actual presentation of the project |
| Evaluate the impact | This is to verify the success of the product based on its performance |
| Final reporting | After all the stages are successfully undertaken, a final report containing all the details is given |

### Milestones

|  |  |  |
| --- | --- | --- |
| milestone | Expected completion | description |
| Data review | Week 1 | Completion of the past data sources |
| Data analysis | Week 2 | Analyzing the initial data acquired. |
| Results of analysis | Week 4 | Give the weaknesses or areas of concerns found. |
| Giving a recommendation | Week 5 | Give a list of requirements that need to be put in place. |
| A preview | Week 7 | Give a report of the current status of the project |
| Develop a working system | Week 9 | Produce the desired system of the project |
| Test the product | Week 10 | Evaluate the working of the system in the real world |
| Outsource the project | Week 12 | Give access to the system to a larger audience |
| evaluate | Week 28 | Evaluate user interaction with the new system |
| Final report | Week 32 | Give a concluding and detailed report |

### Estimated project period

With all the requirements available and the environmental constraints at bay, the project should take me 35 weeks to complete.

## Project Budget

This is the estimates on all the costs from the project,

|  |  |
| --- | --- |
| **ITEM** | **COST(ksh)** |
| Equipment (laptop, implementation, network, coding) | 30,000.00 |
| Related applications | 1,000.00 |
| Printing and binding | 1,000.00 |
| Communication and supervision | 3,000.00 |
| Miscellaneous | 7,000.00 |
| **Total costs** | **42,000** |

## Other materials

Scikit-learn.org

[www.sciencedirect.com](http://www.sciencedirect.com)

sist.sathyabama.ac.in

Journal of survey in fisheries sciences

Paloalto.com