# YOUTUBE SPAM CLASSIFICATION USING N-GRAM

**NURUL AQILAH BINTI ISAHAK**

**UNIVERSITI TEKNIKAL MALAYSIA MELAKA**

# YOUTUBE SPAM CLASSIFICATION USING N-GRAM

NURUL AQILAH BINTI ISAHAK

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Computer Network) with Honours.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2020

## DECLARATION

I hereby declare that this project report entitled

**[YOUTUBE SPAM CLASSIFICATION USING N-GRAM]**

is written by me and is my own effort and that no part has been plagiarized without citations.

STUDENT : _____ Date : _____
(NURUL AQILAH BINTI ISAHAK)

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of Bachelor of Computer Science (Computer Network) with Honours.

SUPERVISOR : _____ Date : _____
MR. NOR AZMAN BIN MAT ARIFF

## DECLARATION

# DEDICATION

To my beloved parents who inspired me,
For their prayers and
all the support they have given and
Always standing beside me.

To my supervisor
For encouraging, motivating and
Believing in me.

## ACKNOWLEDGEMENTS

Alhamdulillah, all praises to Allah S.W.T., The Most Greater and The Most Merciful, because without His permission, I cannot finish my research. I also wish to express our gratitude to my supervisor Mr. Nor Azman bin Mat Ariff for his guidance, invaluable help, encouragement and patience from all aspects until the end of the project. His kindness and suggestion during the preparation of this research are highly appreciated.

In addition, I would also like to express my gratitude to my family, especially my parents. guidance, support, encouragement and advice have given are very helpful. Without them, I would not be able to overcome the problems encountered along the way in my studies. Thank you very much.

# ABSTRACT

Social networking is the phase used for user interaction with others. Spam has become a trend attack and most YouTube users do not know and not aware of spam attacks. Unsolicited bulk mail known as spam has become one of the most Internet's disruptive issues. This research uses the Support Vector Machine (SVM) to develop a YouTube detection framework. The dataset will be used is YouTube spam dataset obtained from the UCI Machine Learning Repository website which is this data most frequently used by the previous researcher. This dataset process through Bag-of-Word, Chi-Square, Information Gain and propose an SVM model that produces from this research. The objective of this project is to prove that SVM can provide result accuracy in spam comments on YouTube. This research is giving hope to produce a system that can detect spam and legitimate comment on YouTube methods based on the SVM model.

# ABSTRAK

Rangkaian sosial adalah fasa yang digunakan untuk oleh pengguna untuk berinteraksi dengan orang lain. Spam telah menjadi serangan trend dan kebanyakan pengguna YouTube tidak mengetahui dan tidak menyedari mengenai serangan spam. Surat pukal yang tidak diminta yang dikenali sebagai spam telah menjadi salah satu masalah yang paling mengganggu Internet. Penyelidikan ini menggunakan Support Vector Machine (SVM) untuk membangunkan rangka kerja pengesanan YouTube. Dataset yang akan digunakan adalah dataset spam YouTube yang diperoleh dari laman web UCI Learning Learning Repository yang merupakan data yang paling sering digunakan oleh penyelidik terdahulu. Proses dataset ini melalui Bag-of-Word, Chi-Square, Information Gain dan mencadangkan model SVM yang menghasilkan dari kajian ini. Objektif projek ini adalah untuk membuktikan bahawa SVM boleh memberikan ketepatan keputusan dalam komen spam di YouTube. Penyelidikan ini memberikan harapan untuk menghasilkan sistem yang dapat mengesan spam dan komen yang sah mengenai kaedah YouTube berdasarkan model SVM.

# TABLE OF CONTENTS

# LIST OF TABLES

**PAGE**

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| **ASCII** | **-** | **American Standard Code For Information Interchange** |
| **BN** | **-** | **Bayesian Network** |
| **BOW** | **-** | **Bag-Of-Word** |
| **CRF** | **-** | **Conditional Random Fields** |
| **CS** | **-** | **Chi-Square** |
| **DFS** | **-** | **Distinguishing Feature Selector** |
| **DFSS** | **-** | **Discriminative Features Selection** |
| **DNS** | **-** | **Domain Name System** |
| **DOS** | **-** | **Denial Of Service Attack** |
| **DT** | **-** | **Decision Tree** |
| **ELM** | **-** | **Extreme Learning Machine** |
| **Email** | **-** | **Electronic Mail** |
| **ERP** | **-** | **Enterprise Resource Planning** |
| **FN** | **-** | **False Negatives** |
| **FP** | **-** | **False Positives** |
| **FS** | **-** | **Feature Selection** |
| **FYP** | **-** | **Final Year Project** |
| **GI** | **-** | **Gini Index** |
| **GIF** | **-** | **Graphic Interchange Format** |
| **GM** | **-** | **Graphic Models** |
| **HMM** | **-** | **Hidden Markov Model** |
| **HTML** | **-** | **Hypertext Markup Language** |
| **IDEA** | **-** | **Integrated Drive Electronics** |
| **IG** | **-** | **Information Gain** |

| | | |
|---|---|---|
| **IR** | **-** | **Information Retrieval** |
| **ISA** | **-** | **Internet Security Attack** |
| **JIT** | **-** | **Just In Time** |
| **KNN** | **-** | **K- Nearest Neighbor** |
| **LS** | **-** | **Lexical Simplification** |
| **MDL** | **-** | **Minimum Description Length** |
| **MyCERT** | | **Malaysia Computer Emergency Response Team** |
| **NB** | **-** | **Naive Bayes** |
| **NLP** | **-** | **Natural Language Processing** |
| **NN** | **-** | **Nearest Neighbor** |
| **PO** | **-** | **Project Objective** |
| **PQ** | **-** | **Project Question** |
| **PS** | **-** | **Problem Statement** |
| **RBF** | **-** | **Radial Basis Function** |
| **RDC** | **-** | **Relative Discriminative Criterion** |
| **SMS** | **-** | **Short Message Service** |
| **SVM** | **-** | **Support Vector Machine** |
| **TF** | **-** | **Term Frequency** |
| **TF-IDF** | **-** | **Term Frequency-Inverse Document Frequency** |
| **TN** | **-** | **True Negatives** |
| **TP** | **-** | **True Positives** |
| **TS** | **-** | **Text Simplification** |
| **UCI** | **-** | **UC Irvine** |
| **Web** | **-** | **World Wide Web** |
| **Weka** | **-** | **Waikato Environment For Knowledge Analysis** |

# LIST OF ATTACHMENTS

# CHAPTER 1: INTRODUCTION

## 1.1 Introduction

The Internet's popularity has affected the conceptual visualization of activities in our life. This influence has become more apparent with the advent of social media, fast and more available Internet and smart technologies. The transition from the physical world to the digital domain has undeniably improved the quality of life significantly, taking into account the ease and time savings offered (Bendovschi, 2015). Individuals, particularly young people, probably don't read e-books, long web pages, and blogs, but rather to read short texts or watch videos. The social network is any web site that provides a network of people within an area to communicate. The most popular social networks are YouTube, Facebook, Instagram, Twitter and others For example, social media platforms provide effective matching of such as short posts, videos, tweets, and even comments.

However, YouTube has become one of the most popular and used social media platforms and is mainly based on the concept of video sharing (Alper Kursat Uysal, 2018). YouTube is currently the world's wide user-driven video content provider and it has become a larger platform for spreading multimedia information (Wattenhofer, Wattenhofer, & Zhu, 2012).Users of YouTube are known as channel and YouTube allows users to upload, rate ,share ,comment on video, subscribe to other user and other feature. However, due to the popularity of YouTube, it has been target of malicious attack.

.

## 1.2 Problem Statement (PS)

The social networking is the phase used for user interaction with other. The rapid growth of social networks has led to various malicious operations. There are various types of malicious attack over the social network which are spam, video spam, malicious link, fraudulent reviews and fake friends or subscribe (Yusof & Sadoon, 2017). Spam has become a trend attack and most of the YouTube users are do not know and not aware of spam attack. Unsolicited bulk mail known as spam has become one of the most Internet's disruptive issues Spam can be classified as unsafe because it has the capability of cyber security threats to end users. The spammer used this chance to spread malware via the comment fields so that it will exploit vulnerabilities inside the user's device (Aziz, Foozy, Shamala, & Suradi, 2018). YouTube is one of the main websites for Internet users to obtain information and spam is now on a YouTube phenomenon attack.

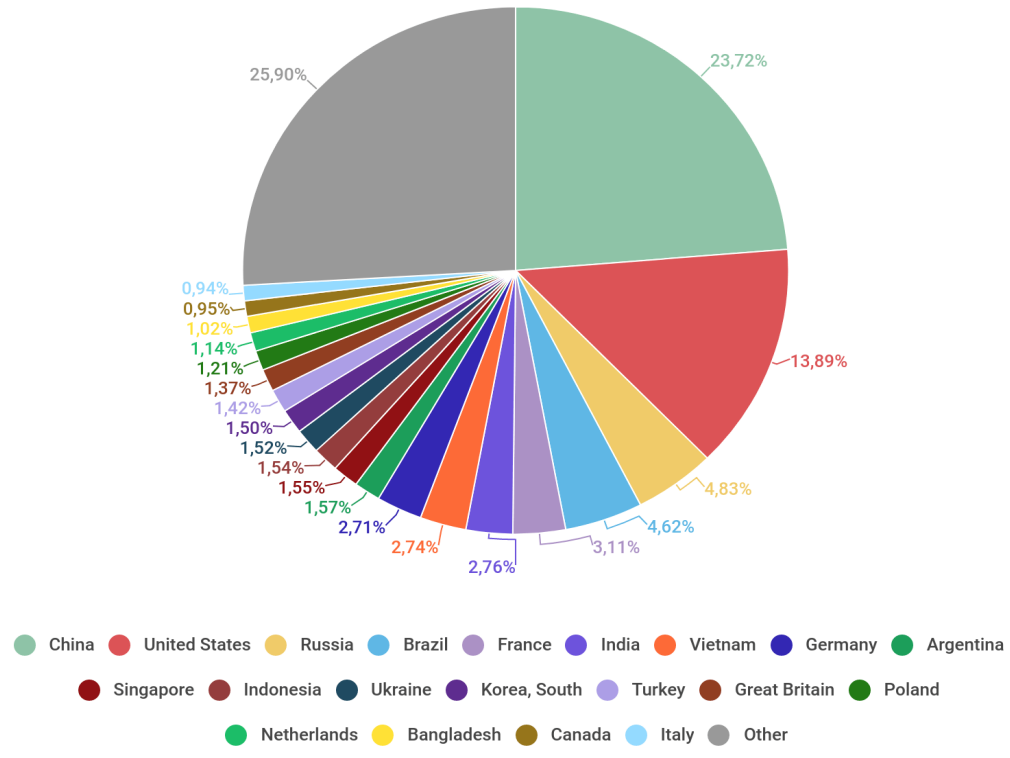



**Figure 1.1: The sources of spam by country, Q2 2019 (SECURELIST)**

According to SECURELIST, as show in figure 1-1 spam and phishing in Q2 2019 has reported that they have got discovered China is the highest source of spam with 23.72 percent leading the list of spam countries, the second position is United Stated (US) with 13.89 percent, whereas Russia at third place with 4.83 percent and followed by other countries. Spam was seen as the fastest growing threat that has infected most electronic devices worldwide (SECURELIST, 2019).

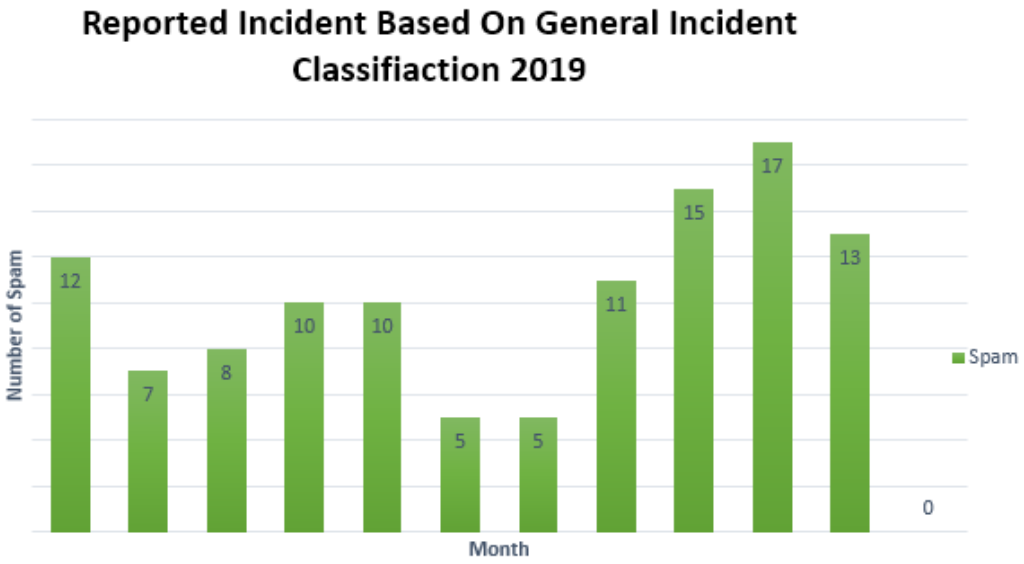



**Figure 1.2: Reported Total Spam by Months (2019)**

**Source: MyCERT Incident Statistics,2019**

Figure 1-2 show the reported incidents statistic by type of crime in 2019. This is a statistical report from the Malaysia Computer Emergency Response Team or MyCERT, provides a reference point for the Internet community in Malaysia to solve computer security incidents. In addition, MyCERT provides services in dealing with incidents such as intrusion, identity theft, malware infection, cyber harassment, and other computer-related incidents. Statistical report shows an accumulation in spam between January and May (first quarter and second quarter 2019) between 12 and 12. Between June to October, we can see the dramatic increase of 5 to 17 collecting. The total number of spam that occurs in 2019, from month January to November is 113

and this report presents only in Malaysia. As a trend, the numbers are expected to grow steadily (MyCERT, 2019).

Over the past two years, another alarming statistics was that the spam attack had fluctuated as the trend as shown in figure 1-3 and figure 1-4 where the total spam for the year 2018 and 2017 are 342 and 344 as reported. As we can look at the trends over time, the spam attack remained constant with 2 events just marginally reduced.

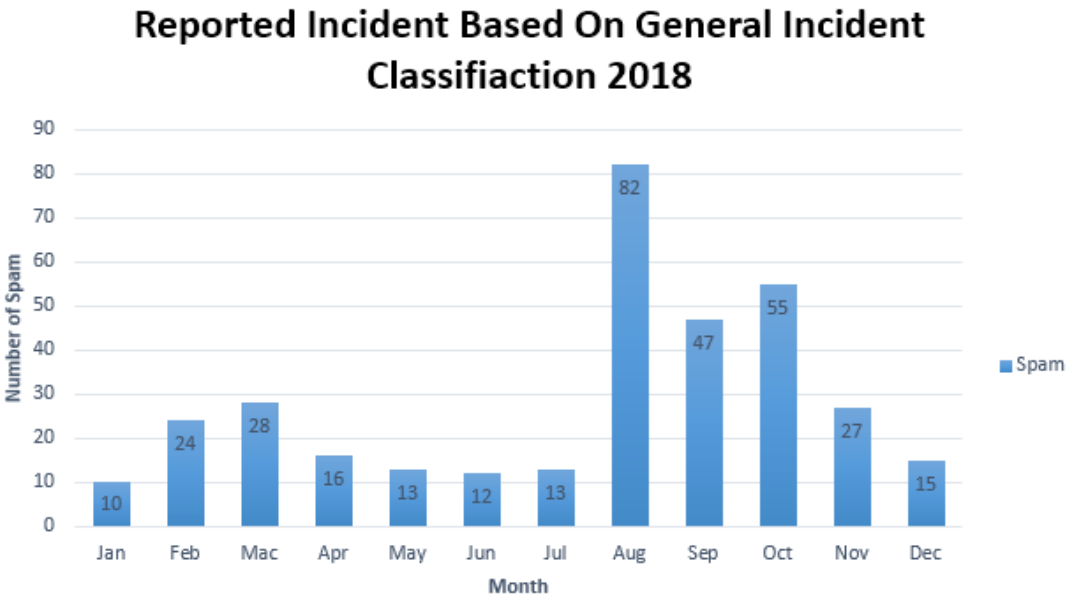



**Figure 1.3: Reported Total Spam by Months (2018)**

**Source: MyCERT Incident Statistics,2018**

In 2018, statistical report shows an accumulation in spam between January and May (first quarter and second quarter 2018) between 10 and 13. Between June to August, we can see the dramatic increase of 12 to 82 collecting. The total number of spam occurs in 2018 is 342. The number of incident reports increases steadily each month (MyCERT, 2018).

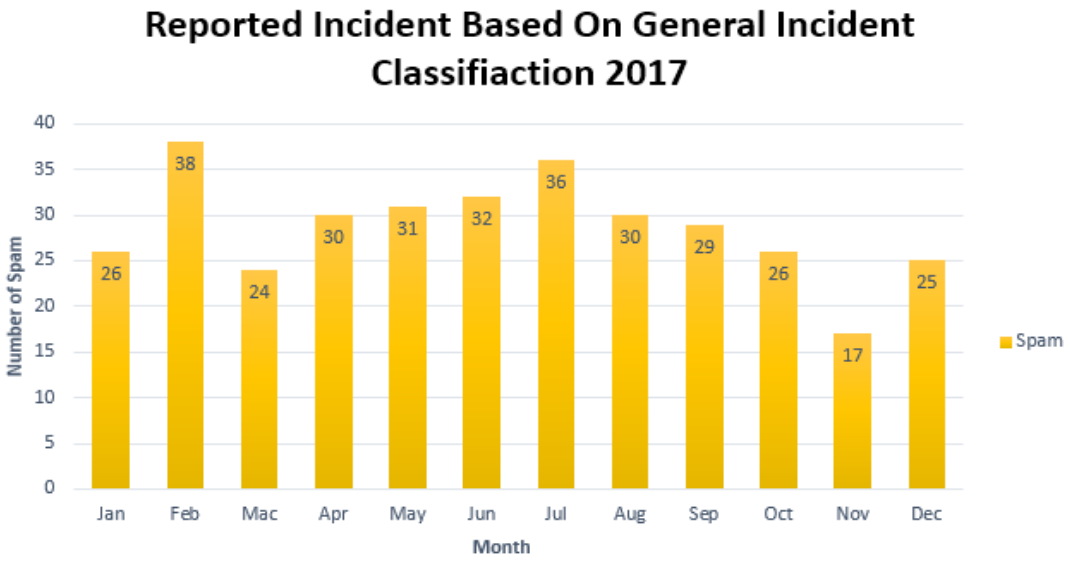



**Figure 1.4: Reported Total Spam by Months (2017)**

**Source: MyCERT Incident Statistics,2017**

In 2017, statistical report shows the number of incidents reported was 344. The accumulation spam between January and May (first quarter and second quarter of 2017) between 26 and 31.We can see the fluctuation between June to December (third quarter and fourth quarter 2017) is 32 and 25 collecting. The highest attack on that year (MyCERT, 2017).

YouTube is another alarming issue with a social media attack. According to Google(2019), YouTube has Community guidelines that set the rules for what does not allow in it. The number of comments removed by YouTube in violation of the community guidelines and filtered as potential spam not accepted by creators in January until March is 228,338,026. The 99.3% ( 226,724,106) of remove comment is detected by automated flagging system and 0.7% ( 1,613,920) by human flagging. The number of comments posted by YouTube slightly increased from April to June 2019 of which 537,759,344 comments were removed. The 99.3% (533,733,538) of remove comment is detected by automated flagging system and 0.7% ( 4,025,806) by human flagging. In Julai to September the number of comments remove decrase to 516,887,894 compared to before this. The 99.5% (514,519,077) of remove comment is detected by automated flagging system and 0.5% ( 2,368,817) by human flagging.

The comments are removed because it contains spam, scam, children safety, cyberbully, and etc.

As we can see, due to the statistics for this year from January until September, it is important that have an effective method be used to predict and detect such comments. Spam detecting on YouTube is considered one of the best ways to solve the problem. The problem was defined in the table below.

**Table 1.1 Problem Statement**

| PS | Problem Statement |
|---|---|
| **PS1** | Increased number of comments on youtube will cause difficulty in distinguishing legitimate comments with spam which causes users to be vulnerable to this attack. In addition, the adversity in getting an effective model of training data in detection of the accuracy of youtube spam on comments. |

## 1.3 Project Questions (PQ)

This problem of the project refers to the problem statement found in this research. The questions you have raised are more about seeking to understand the study you want to do. In this research, there were three project questions that required answers before this research.

I. What is spam comment on YouTube in social media?

II. How YouTube spam is classified?

III. Does the machine learning technique detect YouTube legitimate comment and spam comment more accurate?

**Table 1.2 Summary of Project Question**

| Problem Statement | Project Question |
|---|---|
| PS1 | PQ1 |
| | PQ2 |
| | PQ3 |

This project will be developed on the basis of the above questions. These questions are important to the successful development of the project.

## 1.4 Project Objective (PO)

Project Objective provide the goals for this research. The research's objectives are explained as follow:

I. To study the taxonomy of YouTube spam.

II. To develop a Machine Learning system that is able to detect comment spam on YouTube.

III. To test and verify the functionality of the tools created to detect comment spam.

Table 1-3 shows the relation of each PS, PQ and PO in this research.

**Table 1.3 Summary of Project Objective**

| Problem Statement | Problem question | Project Objective |
|---|---|---|
| PS1 | PQ1 | PO1 |
| | PQ2, PQ3 | PO2 |
| | PQ1, PQ2, PQ3 | PO3 |

## 1.5 Project Scope

The scope of this project is limited to the following criteria shown below:

I. Dataset collected from UCI Machine Learning Repository.

II. The project only focus on Machine Learning method only.

III. Focus on YouTube spam only and only process of content in a comment.

## 1.6 Project Contribution

By project implementation, evaluation properties depend on spam comment in YouTube detection creation. The summary of this project contribution is shown below:

I. Identification of various YouTube spam comments attack based on the taxonomy of its attributes

II. Propose a model that is able to detect YouTube spam comments attacks

III. Development Tool that propose will done with test and validate by verify rate of accuracy in YouTube spam detection.

## 1.7 Thesis Organization

The report will be complete with six chapters which are begin with Chapter 1: Introduction, Chapter 2: Literature Review, Chapter 3: Project Methodology, Chapter 4: Analysis and Design, Chapter 5: Implementation, Chapter 6: Testing and Validation, and done with Chapter 7: Project Conclusion.

### 1.7.1 Chapter I: Introduction

This section concentrates on introduction, project background, project statement definition, project issues, project goals, project scope, project contributions of the studies.

### 1.7.2 Chapter II: Literature Review

This chapter include earlier task and experiments that can be used by previous researcher and a critical review on this chapter that will be explained in this project..

### 1.7.3 Chapter III: Project Methodology

This section touches mostly on what methodology for this project method would be used to evaluate. This chapter would also justify its methodology of each advancement used and the milestone of the project

### 1.7.4 Chapter IV: Analysis and Design

This section may concentrate on analysing the entire debate around this project assessment with architecture that was used for this project with the assistance of the necessity for hardware and software along the course of this project.

### 1.7.5 Chapter V: Implementation

In chapter 5, to get the outcome of precision, the test method will be explained along. The outcome of this experiment will be gathered to prove that this experiment is complete and that the outcome will be registered to create an assessment and then compared to other methods. Identify software environment setup, software configuration management and define the execution of each process that will be created will be included in the project operation.

### 1.7.6 Chapter VI: Discussion

This section describes the outcomes of the past chapter. Summarize and clarify the results obtained from the implementation. Other than that, it will clarify how the outcomes are being analysed from all perspectives and how they are being achieved.

### 1.7.7 Chapter VII: Project Conclusion

In addition, this section will clarify all project summary, project contribution, and project restriction. In this chapter, all the procedures that are used implemented will be shortly listed. Finally, this section will clarify any future work that may be accomplished in the future as well.

## 1.8 Conclusion

In conclusion, this research is conducted to improve knowledge about detect spam comment attacks and gain more feature to improve the prediction process to identify legitimate comments or spam comment. Additionally, this investigation will generate a framework for detecting and preventing YouTube spam comment attacks to help users and tools for tracking events using specific methods of machine learning. The following chapter will discusses the relevant works or literature based on internet security attack and machine learning in detail.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Introduction

This chapter discuss and review the earlier research on the internet security attack using classification methods as show in Figure 2.1 to develop a new models. There should be a better understanding and awareness of spam attack at the end of this chapter. This chapter included the providing literary works for evaluation from prior studies or study.The literature review will be divided into several sections to describe internet security attacks spam, machine learning, the selection of features to be used, the division technique and classification used by researchers during the literature review.

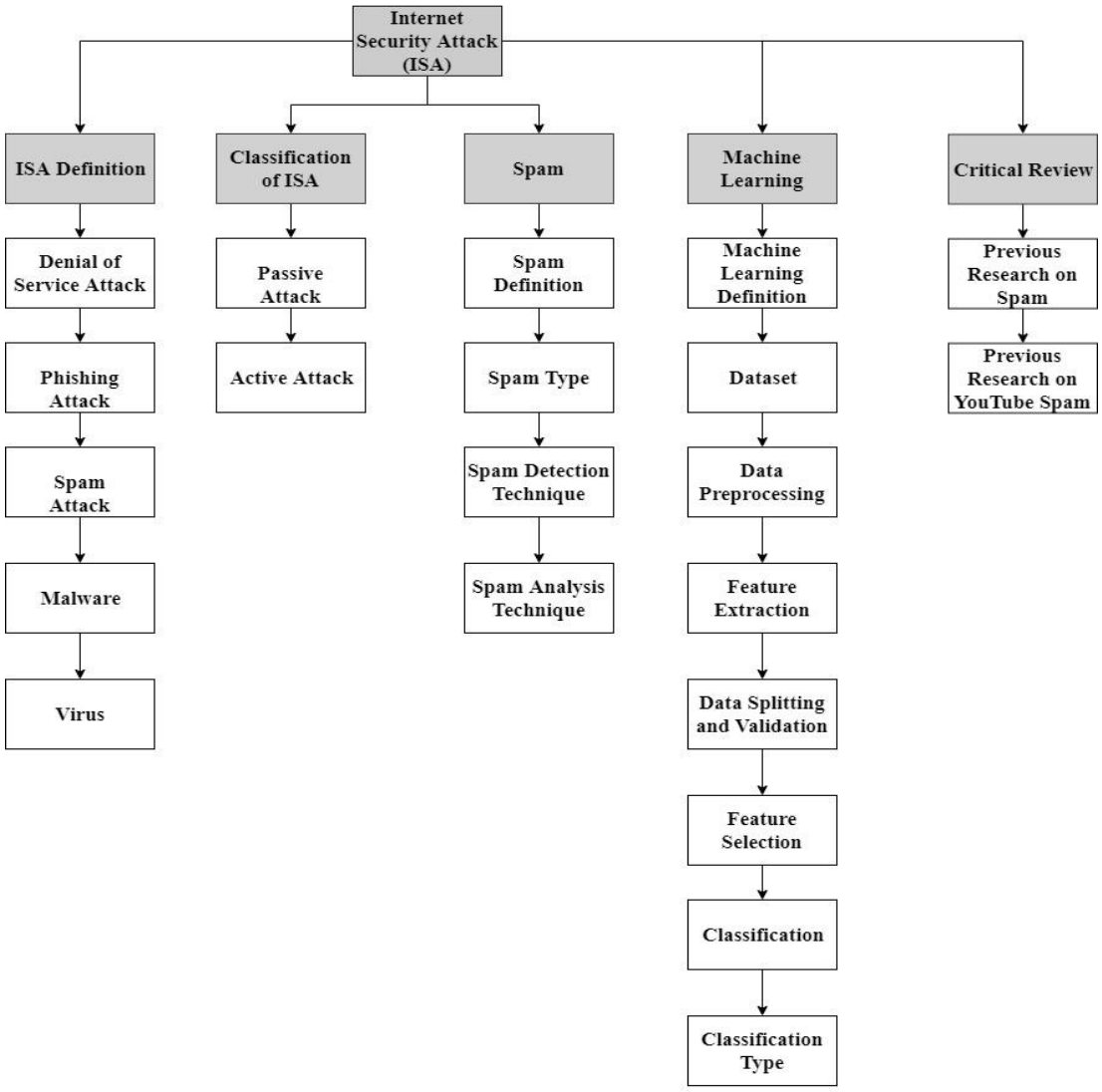



**Figure 2.1: Literature Review's Structure**

## 2.2 General Categories of Internet Security Attack (ISA)

### 2.2.1 ISA Definition

Internet is a collection of several computers around the world that are connected with each other, it is a platform which does not limit the information of every user (Puspita & Rohedi, 2018).The growth of the Internet led to malicious attack. Attacks within the network is attempting to steal, manipulate, destroy or gain unauthorized access or use of property (S.Mangrulkar, R. Bhagat Patil, & S. Pande, 2014). Internet security is the main issue of computing, as many types of attacks are increasing every day and malicious nodes create a network problem (Pawar & Anuradha, 2015). According to Zhang and Gupta (2018), the security attack is a threat used by unauthorized users to steal and modify data from the system.

### 2.2.2 Denial of Service Attack (DoS)

DoS attacks attempt to interfere with a host or network resource to block authorized users from accessing the computer system (Aawadallah 2015). An unauthorized user tries to send many authentication requests to a server using a random user ID at a one time to congest the network and reject other authorized users to use the network services (Tiwari, 2014). The type of threat that degrades the network so that resources are not available to authorized users is known as Dos attack. Malicious hackers are flooding attack computer with inefficient traffic so it could fail to respond or respond slowly and often are inaccessible. There several examples of DoS attack which are Ping of death, SSPing, Land, Win Nuke and SYN flood (Ahmad & Habib, 2010).

### 2.2.3 Phishing Attack

Phishing is a technique aimed at stealing users personal information with disguise as a reliable source (Bendovschi 2015). Phishing is an attack by other person, commonly for financial gain, to request confidential information from a user in individual, group, or organization (Konakalla & Veeranki, 2013). Phishing attacks in which an attacker attempts to obtain personal data from a target user through a fake webpage that exactly matches the existing site (Z. Zhang & Gupta, 2018). The part of the phishing attack was the scammers would then send out an email message trying to trick the user into clicking a link that led to the fake site. When a user tries to sign in with their account details, the hacker records their username and password and then attempts information on the actual website. (Akanksha, 2014)

#### 2.2.3.1 Clone Phishing

Clone phishing is a type of phishing attack in which hackers try to clone a website that is commonly visited by users (M.Nazreen & S.Munawara, 2013). This is a type of attack in which a valid email containing an attachment or the URL has been used to create a cloned email with its content address and recipient (Khan, 2013). Clone phishing email that was once sent is used for cloning malicious emails. Malicious emails usually contain links to phisher websites(Chaudhry, Chaudhry, & Rittenhouse, 2016). Attachments or links in emails are replaced with malicious versions and sent from fake email addresses to the original sender(Khan, 2013). Phisher sets the email settings by placing his email address in the reply section. Therefore the message will be answered to the phisher instead of the user-self when a user sends an email. Phishers then cloning the answer and send it to the user. (Ma, 2013)

#### 2.2.3.2 Spear Phishing

Spear phishing is a technique that aims to target a specific victim. Before the attack, the target details were known and the email was sent from a known source (Khan 2013). Spear phishing is an email sent to specific users who believe it was from a recognized person or group. According to M.Nazreen and S.Munawara (2013), spear phishers target a select group of people who have something in common, not by sending thousands of messages at random. Spear phishing attacks are usually preceded by a setting up and researching the potential victims. An attacker can then send a message from a valid source(Chaudhry et al., 2016).

#### 2.2.3.3 DNS Base Phishing

DNS-based Phishing or pharming modifies host files that are used for Domain Name System (DNS) exploitation. As a search for Links or name services, it returns fake addresses and take to fake sites and users can also send private information to fake sites (Chaudhry & Rittenhouse, 2015). Pharming is an attack aimed at redirecting traffic from a website to another fake site. Pharming interferes with the domain name transfer to an IP address so that the actual website's domain name is converted to the fake website's IP address (M.Nazreen & S.Munawara, 2013). People do not know this

and will enter personal data and the attackers will hack it and possibly are not in the same country (Suganya, 2016).

### 2.2.4 Spam Attack

The term Spam refers to unsolicited e-mails and related unwanted online communications (J. M. Rao & Reiley, 2012).Spam is the misuse of online messaging services for the sending of unwanted messages in bulk (Hayati et al., 2010). Spam comes in many forms and has many different varieties. This may be a preliminary contact for cybercriminals, such as fraudulent scheme operators using email to get money from victims or for identity theft by forcing recipients to share personal and financial account details (Broadhurst & Alazab, 2017).Spam block the receiver's mailbox and uses essential storage space for mail. Spam also includes inappropriate and offensive material that young people are not allowed to access. Many spam messages promote trafficking goods, pirated technology and criminal activity (Krishnamurthy, 2015).

### 2.2.5 Malware

Malware attacks are the main threat in networks. Malware is a malicious code that contains worms, viruses, bots, and spyware. Self-propagation is a basic feature of malware, such as malware that can exploit vulnerable hosts and use affected hosts to spread it.(Z. Chen & Ji, 2009).Malware continues to improve its quantity and sophistication solutions with the addition of invisible samples or types of malware that also provide sufficient diagnostic information to address threats with the lowest human load (Rudd, Rozsa, Günther, & Boult, 2017). Based on Kondakci (2008), malware or malicious code is a software program that can be aggressive, disruptive or irritating and also some forms of malware do not recreate it. Malicious software or malware already exists on mobile platforms, but its frequency and severity are limited. (Mylonas, Dritsas, Tsoumas, & Gritzalis, 2012)

### 2.2.6 Virus

Computer viruses and worms are a growing issue among computer users worldwide. The phrase virus refers to malware that requires the help of a computer user to transmit to another device. According to Balthrop et al. (2004), email viruses allow someone to view a message or open the attached file to spread it. Computer infection like viruses and worms spread through computer-to-computer networks, with various types of networks being infected by different types of viruses. Computer viruses are programs that can affect other programs by modifying them to provide a copy that has been created. The virus will spread over a computer system or system with every user's permission to infect the software (Subramanya & Lakshminarasmhan, 2001).Viruses were once distributed by exchanging disks, but now global networking makes it possible to spread malicious code quickly (Mishra & Ansari, 2012).

## 2.3 Classification of ISA

Network attacks can be classified into two types which are passive and active attacks. A passive attack when a network attacker intercepts information over a network and an attack is active when an attacker initiates an order to interrupt the normal operation of the network(Pawar & Anuradha, 2015).

### 2.3.1 Passive Attack

According to Ramesh (2013), a passive attack compromises the system's privacy and is hard to detect because it gets the information needs without modifying system resources. This type of attack requires the use of system data but does not impact system resources. The intruder will monitor the transmitted data or information between the sender and the recipient. Passive attacks where data is gained by the attacker and do not intend to change the original message content. It is very hard to identify because it does not modify the information (Ahmed, Verma, Kumar, & Shekbar, 2011). The passive attacks are also known as traffic analysis, Eavesdropping, and Monitoring (Pawar & Anuradha, 2015). Phishing is a kind of passive attack (Suganya, 2016).

### 2.3.2 Active Attack

Based on Ahmed et al. (2011), an active attacks are attacks that change the original message or create a false message. These attacks are very complicated and cannot be easily prevented. An attacker or intruder hacks into the network and changes the capabilities of the system. This threat tries to break protective features, implement malicious code and obtain information. Active attack or threaten the credibility of accessibility as it changed the operating system and restrict access to potential users (Ramesh, 2013). It can classify into 3 types which are interruption, fabrication, and modification (Ahmed et al., 2011). Active attack types include spoofing attacks, Wormhole attacks, Modification, Disclaimer, Sinkhole, and Sybil (Pawar & Anuradha, 2015).

## 2.4 Spam

### 2.4.1 Spam Definition

In the medium of communication, spam or unsolicited email is disruption and be within the sort of advertising or similar specific content which can contain malicious code hidden in it (Trivedi, 2016). Spam became one of the most Internet's irritating issues and it was a huge problem for most users as it clutters their mailboxes and spends their time deleting spam until reading the appropriate ones (Issac 2010). The most recognized type of spam is e-mail spam, but the phrase spam is used in other media and media to describe a similar offense. According to (Hayati et al., 2010), Spamming is the dissemination of unsolicited and irrelevant content in various domains such as email, instant messaging, web pages, IP telephony, etc. The unwanted bulk e-mail is become seriously, digital messages sent randomly to thousands of recipients. Spam features change over time, which allows the rules to be implemented (Androutsopoulos, Koutsias, Chandrinos, Paliouras, & Spyropoulos, 2000).

### 2.4.2 Spam Type

#### 2.4.2.1 E-mail

The email has been recognized as the most effective and easy medium of communication since Internet users have rapidly increased. Email spam generally is an electronic message that is spam if the recipients ' identity and background are irrelevant because the email is equally applicable to many other possible recipients; and the recipient has not granted deliberate, clear, and still-revocable permission to send it verifiably (Christina, Karpagavalli, & Suganya, 2010). Email spam has become one of the Internet's biggest problems, causing financial harm to businesses and disrupting individual users. Based on Issac (2010), the process of sending unwanted e-mail messages to e-mail users is e-mail spam, classified as junk e-mail, unsolicited bulk e-mail or unsolicited commercial e-mail. Mail spam refers to send to the multiple people insignificant, unwanted and unsolicited email messages. E-mail spam is intended to advertise, popularize and spread backdoors or malicious programs (Hayati & Potdar, 2008).

#### 2.4.2.2 Web Spam

The web search engine has been an effective tool for searching on the Internet for the information needs of users. The number of web spams is gradually increasing, which has significantly affected the web browser user experience (Li, Nie, & Huang, 2018). Web Spam is a wide known phenomenon to all users that browse the Internet on a daily basis through a search engine. It is a frustrating experience as it requires users to load pages whose content is often completely unrelated to the search engine request that has submitted.(Becchetti, Castillo, Donato, Leonardi, & Baeza-Yates, 2008). Refers to (Hans, Ahuja, & K. Muttoo, 2014), spammers have successfully developed new sophisticated spam spreading techniques. The major reasons are revenue generation, higher search engine ranking, goods & services promotion, data theft, and phishing.

There are three types of web spam which are link spam, content spam and cloaking, a technique in which the content presented to the user's browser is different from the content presented to the search engine..(Araujo & Martinez-Romo, 2010). For the link spam is manipulating the structure of links or anchoring text between

pages for a higher rank (Hans et al., 2014). Content spam refers to changes in the pages content, for example by embedding a large number of keywords. The spam content is potentially the first and most popular type of web spam (Becchetti et al., 2008).And Cloaking is the way to provide crawlers and users with different versions of a site based on information found in a request (Spirin & Han, 2012).

#### 2.4.2.3 Short Message Service (SMS) Spam

SMS usually sent to a wide number of recipients. SMS spam has a significant economic effect on end-users and service providers.(Mahmoud, Mahfouz, Minia, & Minia, 2012). SMS spam has a greater effect on users than email spam as users look at each and every SMS they receive, and users are directly influenced by SMS spam. There are two big approaches which are collaborative and content-based methods for identifying SMS spam. The first is based on user feedback and shared user experience, while the second is focused on analyzing messages textual content(Karami & Zhou, 2014). According to (Abdulhamid et al., 2017), because of its pervasive nature, SMS spamming has become a big nuisance for mobile subscribers and it involves considerable costs in terms of lost productivity, network bandwidth usage, management and personal privacy raid. Mobile SMS spams irritate smartphone users and cause new social stress to mobile phone devices just like e-mail spam.

#### 2.4.2.4 Image Spam

Image spam is just like email spam where spam message text is displayed as a photo in the image file, spammers used it as a way to avoid text-based spam filters (Annadatha & Stamp, 2018). Image spam is the latest trap created by spammers and is a pretty effective way to cheat spam filters as they can only detect text. An image spam email can be described as a document formatted with HTML, which is typically an un-suspicious text and an encoded image sent as an attachment (Das & Prasad, 2014).The message text is now commonly named photo spam, mostly taken from internet sources, such as news articles and message boards. According to (Dredze, Gevaryahu, & Elias-Bachrach, 2007), image spam allowed spammers to form spam as an anti-spam captchas.

#### 2.4.2.5 YouTube Spam

YouTube has become one of the social networks most visited and used. The platform is based primarily on the concept of video sharing. YouTube is one of the largest sites on the Internet for users to obtain information. That's why many spammers are going to trick YouTube users by spamming the comments on YouTube. The spammer used this chance to spread malware across comment fields, exploiting vulnerabilities in the computers of the user (Aziz et al., 2018). Spam is now a trend attack and spam is described by YouTube as inappropriate comments, such as harassment or trolling, and people trying to sell things. Based on Hamou & Amine (2013), spam is now a trend attack and spam is described by YouTube as inappropriate comments, such as harassment or trolling, and people trying to sell things. Spam causes many problems, such as wasting time, energy, and using bandwidths from the network.

Because of the risk of spam, organizations and consumers may face financial loss. Some of the spammers use the comment section on YouTube for ads, while others are responsible for spreading computer viruses and some spam messages are intended to steal financial identities from the user (Samsudin et al., 2019). Spam's more concerned threats are when malicious spam is involved, which will result in phishing websites when users click on the link and malware distribution. A large YouTube user base is made up of children who are often exposed in the form of comments to malicious and harmful material (Aiyar & Shetty, 2018). YouTube site has been overwhelmed by the content of very low quality that can be considered as spam videos and spam comments. Spam comments can be described as comments that are not relevant to specific content, announce sales, promote pornographic content, undermine the credibility of the website, and make the website more reliable by increasing the number of views (Abd & Qaisar, 2019).

### 2.4.3 Spam Detection Technique

Image spam is the latest trap created by spammers and is a pretty effective way to cheat spam filters as they can only detect text**.** Image detection of spam can be based on content. That is to say, first try to determine the text in a spam image and then make filtering decisions in the same way as with non-image spam (Attar, Rad, & Atani, 2013). The different techniques for spam detection of images show spam detection.

Based on low-level image features (visual features) consist of color, texture, shape, and appearance features. The others technique uses near duplicate detection in images. It includes the clustering of Gaussian Mixture Models (GMM) based on the principle of Agglomerative Information Bottleneck (AIB) (Ketari, Chandra, & Khanum, 2012). Next use high-level image features provide image header information such as file name, file format, aspect ratio, area of image, compression, horizontal, vertical resolution, etc. It used four basic image features which are width, height, type of image file and file size, which can be easily obtained via an image at an extremely low computational cost (Attar et al., 2013).

A variety of techniques for spam detection exploit the features found in text-based sites. In addition, users of such systems can easily learn to identify, skip or avoid such text spams, such as URLs. (Chowdury, Monsur Adnan, Mahmud, & Rahman, 2013). Official blog on YouTube reported efforts to resolve unwanted comments by identifying malicious links, detecting ASCII art, and showing changes to long comments. The word-based representations of the text include a tokenizer to divide the message into tokens and typically a lemmatizer to minimize the set of words (Kanaris, Kanaris, & Stamatatos, 2006). You can use many detection techniques to find YouTube spam comments. The detection technique is used to detect spam comments and hum comments. Spam detection uses the bag of words representation, which means that each message is considered to be a set of words that occur a number of times. As mentioned, this research will be done by relying on the text. Spam detection algorithms can be classified as origin based technique and content based technique.

For origin based spam detection technique it is divided into three part which are whitelist, blacklist and realtime blackhole list (RBL). The white list is a list containing all the emails we always want to receive mail from. An interesting choice is an automated whitelist system that removes the need for administrators to manually enter approved addresses on the whitelist and guarantees that mail from unique senders or domains is never classified as spam. Whitelist filters will not embrace e-mails from any address other than the list of good e-mail addresses (Issac, 2010). Furthermore, the same as a competitive alternative, blacklist list serves as a list of addresses that we do not want to accept it (Christina et al., 2010). Blacklist filters permit emails from any

address other than the list of known bad email addresses. The blacklists can be maintained and conducted locally or accessed via the Internet. Then the realtime blackhole list or RBL are also known as the blacklist list available on the Internet. Sometimes people compile the real-time blacklists on the Internet and put whole ranges of IP addresses on their blacklist, even though previous abuses occurred only for a certain part of that range (Reddy & Reddy, 2019).

The content based spam detection technique it is divided into four part which are rule based filter, bayesan filter, support vector machine and artificial neural network. The rule based filters are based on some of the filtering guidelines that are used to determine if incoming messages are spam. Rules for e-mail, message content, keywords and e-mail header information can be set (Jiansheng & Tao, 2008). Next, this method is based on Bayesian filters using the Bayes formula. These filters must be trained from a set of recognized good and bad e-mails when implemented as software (Alberto, Lochter, & Almeida, 2016). Furthermore, the support vector machine (SVM) are computationally powerful tools for supervised learning, widely used in problems of classification, clustering and regression. SVM is a strong statistical learning theory used frequently for classification of data. SVM in various applications for data mining makes it a compulsory tool for product development that has implications for human society (Nayak, Naik, & Behera, 2015). Lastly, the artificial neural network (ANN) is a set of connected input or output units with a weight associated with each connection. The network learns during the learning phase by adjusting the weights so that the input tuples can predict the correct class label (Kumar Sharma, Kumar Prajapat Assistant Professor, & Aslam, 2014). ANN is a powerful tool used for data classification, it has the ability to better learn large amounts of high-dimensional data (A. S. Rao, Avadhani, & Chaudhuri, 2016).

### 2.4.4 Spam Analysis Technique

The spam analysis technique that uses in image spam by the detection system aims to extract embedded text in combination with the visual attribute such as colour, texture, shape and hence used to measure the object similarity. Spam image appears to have a rough RGB / LAB color space distribution (Annadatha & Stamp, 2018). The anti-spam community litter anti-spam blogs recommend various Adhoc image spam solutions, including blocking all gif images using FuzzyOcr and limiting advanced

image processing to small images. The gradient orientation histograms are used in image processing in computer vision for object detection purposes. The technique counts gradient orientation occurrences in localized portions of an image and can reveal text characteristics (Attar et al., 2013).Other analysis technique is primitive length uses primitive texture lengths in various directions as a description of the texture. A primitive is a constant collection of total pixels that have the same gray rate in the same direction (Mehta, Nangia, Gupta, & Nejdl, 2008).

For the spam analysis technique that uses in detecting text spam by various machine learning techniques that uses by researcher. Machine learning is part of supervised general leaning and, in particular, text classification. A variety of open source text analysis tools, such as information extraction and classification (Chaudhari & Govilkar, 2015). There are several phase to process the spam to determine the text is legitimate or not. For the pre-processing phase, data cleaning such as tokenization, removal of stop words and stemming will be performed by the raw data set. for the next process, the data clean is used for the feature selection and extraction to generate vector and dictionary functions as an input to the algorithm selected (Aziz et al., 2018). Then, in classification phase the classification of text is intended to assign predefined categories to text documents. We identify the test set after training the classifier and compare the estimated labels with the true labels and measure the performance. Precision, Accuracy, Recall and F-measure will be used to perform the test (Allahyari et al., 2017).

## 2.5 Machine Learning

### 2.5.1 Machine Learning Definition

Machine learning is an application that can automatically ace and complete improvement without the need to be customized or programmed needs, depending on past experience.In machine learning, computer programs that approach information will be grown so it can learn without anyone else. Refer to (Tran, 2019), Machine learning is a concept that can refer to learning from past experience, which in this case is previous data to improve future performance, and it requires promoting an advanced level approach by providing the data that is essential for a machine to train and adapt when exposed to new data.

Machine learning is a research field that is formally focused on learning systems and algorithms theory, performance, and properties. The field of machine learning is classified into three sub-domains: supervised learning, unsupervised learning and reinforcement learning (Qiu, Wu, Ding, Xu, & Feng, 2016) refers to figure 2-1. It is fundamental to understand the form of machine learning so that it can experience the right learning condition for some random undertaking and understand why it worked.

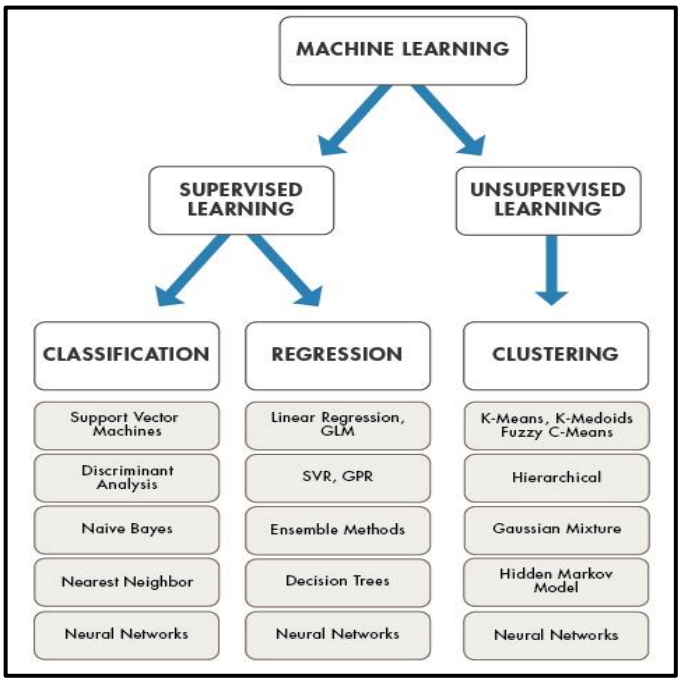



**Figure 2.2: Type of Machine Learning**

.

Supervised learning includes a function from a particular set of data (inputs with the respective outputs). The training data is made up of a series of examples that train the computer (Nath & Dasgupta, 2016). The model is trained using labeled and classified data. It is also sometimes called an intelligent learning model (Tewari & Jangale, 2016). For unsupervised learning, the objective is to achieve a good internal representation of the input. Based on Nath & Dasgupta (2016), in unsupervised learning, labeled examples are not available. The training process is carried out on data that is unlabeled or not classified. The model for the unknown output is to be trained. It requires a deep understanding of mathematical concepts (Tewari & Jangale, 2016).

### 2.5.2 Dataset

The dataset is considered as the cornerstone in conducting research on different goals in a large sample. With this dataset, a new algorithm for comparison using existing algorithms is used as a benchmark (Verbert, Manouselis, Drachsler, & Duval, 2012). Experiments that can overcome the challenge of developing engineering training will use the search for existing datasets. According to Duval (2011), it needs to be better measured obtained, analysed and reported on learner information found by certain researchers in the field of Technology Enhanced Learning (TEL). In this project will use available dataset that focus more on YouTube spam which is a source from the public study repository for UCI, reference by many researchers.

### 2.5.3 Data Preprocessing

#### 2.5.3.1 Definition

Data preprocessing consists of a selection of data attributes, cleaning of data and resolution of missing value. The traditional method of pre-processing data is reactive as it starts with data that is expected to be ready for review and there is no input and data collection system (N. Zhang & Lu, 2007). The data preprocessing is the first step in the process of data preparation, with the aim of reformatting the original logs to identify user sessions. Data Preprocessing is used to clean the data and identifies the technique of discovering the navigation pattern of the user when it provides the pattern to be discovered (Sivakumar, 2017).

#### 2.5.3.2 Preprocessing Type

The image processing technique called pre-processing has been applied to enhance the image quality. Pre-processing of images is one of the initial measures necessary to ensure the high accuracy of the subsequent steps. An enhanced image resolution and pre-processing technique was proposed for the cleaning or removing a noise (Shameena & Jabbar, 2014). Preprocessing is to exploit the signal statistical model as far as possible, thereby reducing the number of algorithm control parameters. Only a significance level is needed to distinguish between signal and noise since all other properties are estimated from the data. (Förstner, 2007). Pre-processing involves conversion, image resize, noise removal and quality enhancement, and creates an

image in which minutiae can be correctly detected. Gray Scale Conversion is the filtering methods that are optimized to noise reduction techniques such as power spectra and a blur filter. Meanwhile, Image resizing is an important part of the image processing technique in order to improve and reduce the image size in pixel format, and image noise is characterized as the random variation of brightness or color information in images generated (Perumal & Velmurugan, 2018).

Preprocessing is an important task in Text Mining, Natural Language Processing (NLP) and Information Retrieval(IR). There are several preprocessing methods for the text documents such as tokenization, stop word removal, stemming and lowercase. Data pre-processing techniques are applied to the target data set to reduce the size of the data set to improve the effectiveness of the IR system (Dr.S.Kannan & Gurusamy, 2015). After reading the input text documents, the text preprocessing stage divides the text document into features called tokenization, Tokenization is generally understood as any kind of pre-processing of natural language text. Tokenization process is applied to the data review, tokenization is the process of dividing the string sequence into single words, keywords, phrases, symbols called tokens (Kadhim, 2018). Tokens are very useful for finding patterns and are considered as a base step for other processing. Most of the words in documents repeat very often but are basically meaningless as they are used to combine words in a sentence. any analysis or generation cannot be carried out (N. Zhang & Lu, 2007).

Stop words commonly not contribute to the context or content of textual documents. Stop words removal are common words such as ’and’, ‘are’, ‘ this’, etc. They are not useful for document classification (Dr.S.Kannan & Gurusamy, 2015). Whereas according to Alper Kursat Uysal & Gunal (2014), some stop words are rare in spam messages and should not be removed from the feature list even though they are semantically void. However, Stop word removal enhanced classification accuracy in most cases and it is proposed that the best choice for text classification is to apply stop word removal (Gharatkar, Ingle, Naik, & Save, 2018).

Stemming methods are used to specify the root or stem of the word. It changes words into their roots, which integrates a lot of linguistic knowledge based on language. For example, the words like user, users, used, using can be stemmed to the

root word 'use' (Gaigole, Patil, & Chaudhari, 2013). It is recommended that stop-word removal and stemming be implemented in order to reduce the dimensionality of the feature space and improve the efficiency of the text classification system (Etaiwi & Naymat, 2017). But there is a risk if it over-stemming or under stemming. Over-stemming appears when two words with different stems are stemmed from the same root. Under-stemming occurs when two words that are not of different stems are stemmed from the same root. In addition, lower conversion is the pre-processing stage for the classification of text. Since upper-case or lower-case word forms are considered to have no difference, all upper-case characters are typically converted to their lower-case forms prior to pre-processing classification tasks (Kadhim, Cheah, & Ahamed, 2015). In both domains and languages, the lowercase conversion is active. In other words, all characters without domain or language dependence should be converted to lowercase. Although lowercase conversion helps to group the words which contain the same details with either upper or lower case characters, fewer features are accomplished with more discrimination (Alper Kursat Uysal & Gunal, 2014).

### 2.5.4 Feature Extraction

Features extraction are selected based on image content features and text content features used for large scale clustering and combined to prevent duplicates. The goal of feature extraction is to examine the benefits of new features in order to achieve high accuracy (A. K. Uysal, Gunal, Ergin, & Gunal, 2013). Feature extraction starts with the initial set of measured data and constructs derived values features intended to be informative and non-redundant. Image spam is rich in content, therefore image spam filters use different methods on their filters to detect image spam aimed at fighting image spammers. There are several feature extraction methods in image processing which are appearance based approach, feature based approach, template based approach and part based approach (Hema & Saravanakumar, 2018). Just in Time (JIT) feature extraction Creates classification time features as required by the classifier. Instead of using complex analysis, this technique focuses on the simple properties of the image and it also uses advanced features such as image size, colour, edge detection and random pixel testing (Ketari et al., 2012).

A lexical comparative analysis used by clients and agents in the Enterprise Resource Planning (ERP) framework and a potential solution for data cleaning and NLP content extraction. It used to identify and extract relevant content that best explains the purpose of the customer and can be used for classification or automated response generation (Nisioi, Bucur, & P.Dinu, 2018). Lexical features might well serve to classify in a cross-lingual task and that semantic characteristics, such as those which can be derived from word embedding and support the identification of paraphrastic reuse (Moritz & Steding, 2019). Lexical features are the character or word-based features, it indicates the types of words and characters the writer likes to use and include features such as the number of characters in the upper case or the average length of the word (Crawford, Khoshgoftaar, Prusa, Richter, & Al Najada, 2015). For the lexical analyzer, pattern recognition engine takes as its input a string of individual letters and divides them into tokens. The lexical analyzer breaks these words into a number of tokens by removing any white space or comments in the code (Chuhan, Singh, Makhan, & Kumar, 2017). Meanwhile, the lexical simplification (LS) is the process of replacing complex words with simpler alternatives of equivalent meaning in a given sentence. The task of LS is to perform Text Simplification (TS) in the context of NLP. In order to make it simple, it can be formally described as the task of replacing words in a given sentence, without any modifications to its syntactic structure (Paetzold & Specia, 2017).

N-gram is a sequence of n items from a general text sequence. N-gram similarity algorithms compare the n-grams in two strings for each character or word (Aiyar & Shetty, 2018). N-gram (n = 1 to 5) statistics and other English language properties were derived for natural language applications and text processing. The algorithm LocalMaxs introduced in order to extract the dominant character n-grams in a corpus. It is an algorithm that calculates local maxima in comparison with similar n-grams for each n-gram. Traditional word bag representation, use the n-grams character bag representation to avoid the sparse data problem that arises in word-level n-grams. The n-grams representation bag of character is language-independent and requires no pre-processing of text (tokenizer, lemmatizer, or other "deep" NLP tools). It has already been used in several projects, including language recognition, authorship assignment and categorization of text based on topics, with remarkable results compared to word based representations. An important feature of the character-level

n-grams is that they prevent the problem of sparse data that arises when using word-level n-grams (Kanaris, Kanaris, Houvardas, & Stamatatos, 2006). Character n-grams can capture nuanced, lexical, syntactic and or structural stylistic details.

### 2.5.5 Data Splitting And Validation

A standard method of calibration with multivariate involves several separate data sets. The validation or training data set is required by estimating an equation's parameters or by training a neural network to set up the model. To estimate the performance of the final model, the validation data set is important. Specific data for all these data sets have been shown, without it the models and predictions are biased. In the research, many researchers use the available data set, which is the dataset, the process will be carried out with training and testing to produce the result. There are several pre-processing methods that have to use 50% training and 50% testing or 80% training and 20% testing (Anguita, Ghelardoni, Ghio, Oneto, & Ridella, 2012). There are various techniques for subsampling, including cross-validation, bootstrapping, and random subsampling.

#### 2.5.5.1 Cross validation

Cross-validation is a statistical method for estimating the ability of models of machine learning. Comparing and selecting a model for a given predictive modeling problem is commonly used in applied machine learning since it is easily understandable, easily implemented, and results in skill estimates that are generally less biased than other methods (Y. Zhang & Yang, 2015). On the concept of data splitting, part of the data is used to match each competing model and the rest of the data is used to measure model predictive performance through validation errors and the best overall performance model is chosen. The data are partitioned into k equal parts for a k-fold cross validation. The first K-fold method is to split the shuffled 100 objects into k chunks of data. Train with (k – 1) chunks of data objects, test them on (k – 1) the chunk and include the one you excluded in the last iteration to train and evaluate your template on this excluded chunk. K-Fold cross-validation is where a given set of data is divided into a K number of sections or folds where at some level each fold is used as a test set. Figure 2.3 shows the example of k-fold cross validation where k=5.

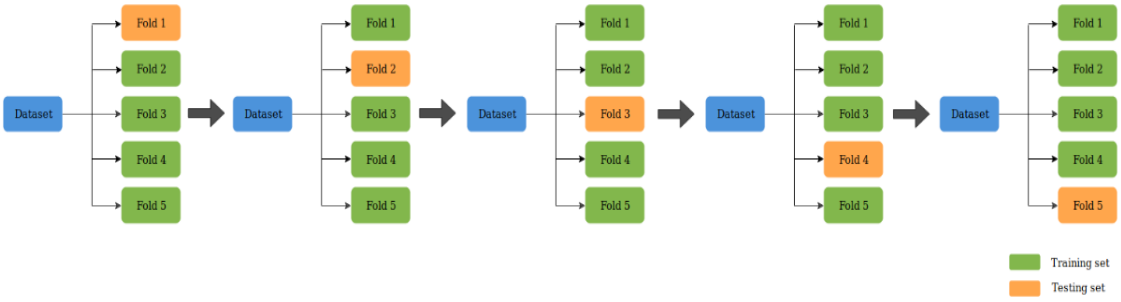



**Figure 2.3 Example of 5-folds cross validation**

### 2.5.5.2 Bootstrapping

The bootstrap approach is a technique of resampling used to predict population statistics by sampling a replacement dataset. Bootstrap resampling was originally developed to help the researcher determine what the outcomes might have changed if they used another random sample instead and how different the expected results might be when implementing a model to new data. Bootstrapping has become very popular in the area of small data sets resampling (Dwivedi, Mallawaarachchi, & Alvarado, 2017). Bootstrapping is based on replacement sampling to create a calibration set. Analysis statistics such as mean or standard deviation can be estimated using it. It is used in applied machine learning to estimate the performance of machine learning models when predicting data that are not included in the training data. We need to choose the sample size and the number of repeats when using the bootstrap. Bootstrapping is not influenced by asymptotic inconsistency and could be the easiest way to estimate the error for very small data sets, whereby the entire process can be replicated arbitrarily frequently.

### 2.5.5.3 Random Subsampling

Random sub-sampling, which is data, is divided randomly into dis-joint training and multiple test sets, and the accuracies obtained from each partition are averaged. The benefit of the random subsampling process is that it can be repeated continuously. Random sub-sampling, also known as cross-validation of Monte Carlo, as multiple holdouts or as a repeated evaluation set, is based on the random splitting of the data into subsets, whereby user determines the size of the subsets (Genuer, Poggi, Tuleau-Malot, & Villa-Vialaneix, 2017). According to previous researcher,

random subsampling is more consistent prediction compare to cross validation and it is suitable to divide the data into subsets of calibration, testing and validation.

### 2.5.6 Feature Selection

#### 2.5.6.1 Feature Selection Definitions

The function is called the feature selection in machine learning, which is generally to improve the efficiency of the data science type in advance. Based on Ramaswami & Bhaskaran(2009), they acknowledges that feature selection in pattern recognition, machine learning, statistics, and data mining communities in a feature selection, were only an active and effective field of the research area. Furthermore, a feature selection also as subset selection, attribute selection and variable selection. t includes the choice of a subset of appropriate features before developing a model. (Wang, Tang, Liu, & Lansing, 2016). Other than that, in the selection of characteristics used by different range data, lower and higher dimensional data were included. It is generally used to extract duplicate and unnecessary information characteristics (Vanaja, 2014). The main goals of feature selection are to avoid overfitting and improve model accuracy and provide better and more cost-effective models (Oreski & Novosel, 2014).

#### 2.5.6.2 Feature Selection Type

Feature selection techniques do not change the original data representation. A feature selection is a common technique in machine learning, learning with a subset of features that can significantly reduce the space of the feature. The several types of feature selection in image processing are average color, color saturation, edge detection, file format, file size, image metadata ,image size, prevalent color coverage and random pixel test (Khawandi, Abdallah, & Ismail, 2019) Feature selection uses the Minimum Description Length (MDL) principle in image processing to perform a sparse selection of models in such a high-dimensional feature space. MDL-based feature selection algorithm. The basic idea for statistical modeling is to represent each model with a distribution of probabilities, based on Kraft inequality, a unique prefix code can be developed for any distribution of probabilities (Liu et al., 2010)

Selecting features will use several types of variable selection methods and can be achieved to clarify the best subset without converting a data into a new set to a reduced value dataset. Other than that, a various-technique variable selection like embedded approach, wrapper method, filter method, chi-square, information gain and hybridity. Machine learning used an embedded approach to variable selection and it throughout the training phase, which should reduce costs and improve efficiency calculation. This approach requires modifications and the evolution of model parameters. As part of the training process, embedded methods include variable selection without splitting the data into training and testing sets (Chandrashekar & Sahin, 2014). Embedded feature selection methods are linked to the classification stage, in this case this link is much stronger as the selection of features in embedded methods is included in the construction of the classifier (Xiao, Dellandrea, Dou, & Chen, 2008).

The filter method is independent of the classification algorithm, evaluate the feature of selected items. Besides, variable ranking techniques are used by filter methods as the main criteria for variable selection by ordering. Ranking methods are used because of their simplicity and for practical applications good success is reported. Ranking methods are filter methods as they are used to filter out the less relevant variables before classification (Chandrashekar & Sahin, 2014). According to Zhu, Ong, & Dash(2007), filter methods use the inherent characteristic of the data to determine the goodness of the function subset. They are relatively cheap because they do not include the algorithm of induction.

Wrapper methods that assess the quality of a subset of features through the performance of a trained classifier. To assess this quality, it requires the use of a classifier (which should be trained on a given subset of features). Based on Chandrashekar & Sahin(2014), the performance of the predictor is the feature selection criterion in wrapper methods, the predictor is wrapped in a search algorithm that will find a subset that gives the highest performance of the predictor. This task would be accomplished by predicting machine learning performance, which features evaluated its efficiency. A wrapper method that is cosine similarity measurement support vector machines (CSMSVM), to remove unnecessary or redundant features during the

construction of the classifier by adding cosine distance to support vector machines (SVM) (G. Chen & Chen, 2015).

The Chi-Square (CS) is a popular feature selection method that evaluates features individually with regard to classes by calculating chi-square statistics. The technique is a separately variable assessment and measurement statistically recoverable (Lee, Lushington, & Visvanathan, 2011). CS feature selection is effectively deployed to identify the optimal feature set that increases model accuracy. CS selection features is a common approach in many applications. According to the researcher (Thaseen & Kumar,2016), by calculating the CS statistic value relative to the class variable, the feature is evaluated in this approach. One of the most efficient methods of feature selection is CS.(Taylor et al., 2015)

Information Gain (IG), this often used data has a high-dimensional classification of text and includes features as counter-number information in parts that will be used for class prediction when an appropriate target is accessible(B. Kumari & Swarnkar, 2011). IG is a selection technique that can reduce the size of features by calculating the value of each attribute and classifying the attributes. IG generally selects the features through scores. The basic idea is that we only have to measure the score for each feature that can be expressed in class discrimination, then the features are sorted according to this score and then just keep those top rankings (Seeja & Zareapoor, 2015)

Hybrid approaches can be used to improve the search algorithm's performance. Through hybridization, to improve the performance of each technique, good properties of at least two techniques are combined (Zorarpacı & Ay, 2016). The hybrid selection scheme that includes both filter and wrapper methods in building a suitable pool of features, coupled with the general lack of success in using individual filter or wrapper methods (Zhu et al., 2007).

### 2.5.7 Classification

There are several machine learning applications, the most important of which is data mining. People are often prone to errors when analyzing or, possibly, when trying to establish relationships between multiple features. This makes it difficult for them to solve some of the problems. Machine learning can often be applied effectively to these issues, improving system efficiency and machine design. Classification is the process in which the class of data points is predicted. Sometimes classes are referred to as target labels or categories (Chaudhary, Kolhe, & Kamal, 2013). Predictive modeling classification is the task of approximating a mapping function from variables of input to discrete variables of output. Classification is a data mining technique (machine learning) used to predict data instances group membership. Although there are several techniques available for machine learning, classification is the most commonly used technique and this is an admired activity in machine learning, particularly in future plans and the exploration of information (Baradwaj, 2012). In addition, classification is a method of data mining which assigns objects to categories or classes. Though classification is a well-known machine learning technique, it suffers from problems such as handling missing data. Missing data set values in both the training and the classification phases may cause problems.(Tilve & Jain, 2017)

### 2.5.8 Classification Type

#### 2.5.8.1 Generative

Generative models that can simulate and forecast sequences of future events will, in theory, learn to capture complex phenomena of the real world, such as physical interactions. Using large unlabeled datasets combined with predictive generative models is an attractive alternative to supervised learning. In order to allow to accurately predict future events, a complex generative model needs to develop an internal representation of the universe (Kumar et al., 2015). Generative models for improving sample complexity and adapting to the shifting distribution of data. By contrast, Generative models are a more natural fit for this form of setup as the maximization of the training target for a new class can be more easily separated from other classes. Previous generative probability as samples may produce artificial data points such as Gaussian, Naïve Bay, Multicultural Mixtures, Hidden Markov Model

(HMM), Bayesians Gaussian and Markov random fields (Yogatama, Dyer, Ling, & Blunsom, 2017).

The Naive Bayes (NB) algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting a given data set's frequency and value combinations. The algorithm uses the theorem of Bayes and claims that all attributes are independent given the category variable value (Saritas, 2019). That is performed by the NB algorithm by assuming conditional independence over the training data set. Despite the simplified assumptions and naivety, NB classifiers work well in complex situations. The advantage of these classifiers is that to estimate the parameters necessary for classification, they require a small number of training data. This is the choice algorithm for categorizing text (Tran, 2019). The classifier of NB is based on the theorem of Bayes and the theorem of total probability.

Bayesian network (BN), also known as belief networks, belong to the probabilistic graphic models (GM) family. Such graphical structures are used to reflect an unknown domain knowledge. Specifically, each node in the graph represents a random variable, whereas the edges between the nodes reflect probabilistic dependence among the random variables (Tran, 2019). It is based on Bayes ' theorem with the doubt of flexibility between each consolidation of features and naive Bayes classifiers function outstanding in certain obvious situations, for example separating spam. The problem is the attributes of conditional density estimation, inference (large, discrete and continuous variables) and multi-dimensional data (Mohammad, 2019). Figure 2-3 shown the structure of bayesian network.

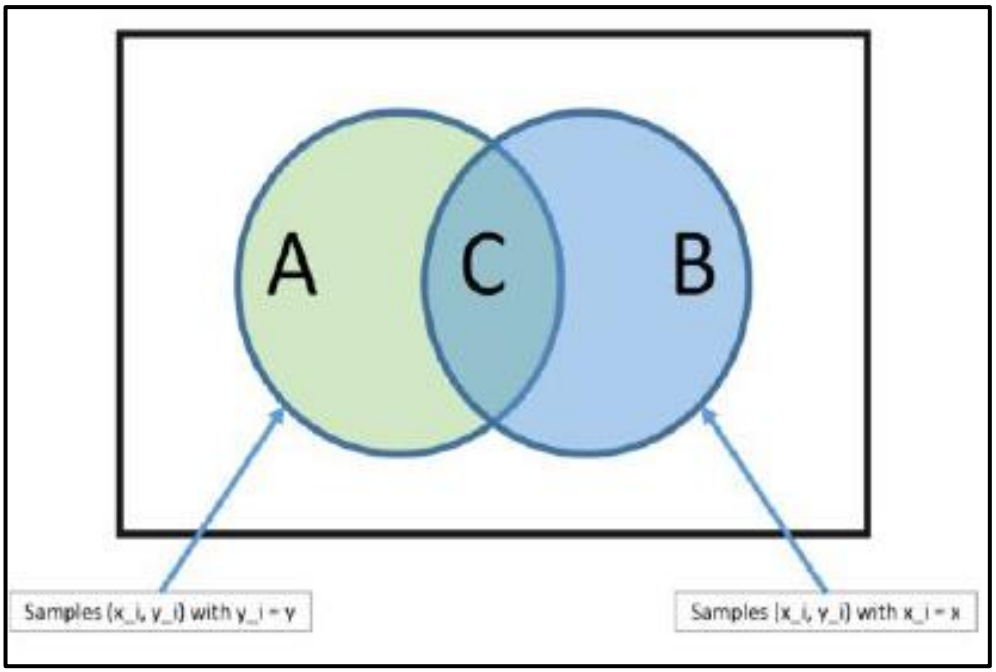



**Figure 2.4: Bayesian network**

Next classifier that researcher mention about Hidden Markov Model (HMM). HMM is a machine learning technique and functions as a state machine. HMM are commonly used to analyze statistical patterns. According to (Nguyen, Khosravi, Creighton, & Nahavandi, 2015), HMM are designed according to a supervised learning approach to enable them to realize available knowledge. The feature of HMMs makes it possible to merge them into larger ones where each HMM is trained separately for each data class. HMM which can handle bag-of-word sequences and each state in the HMM has a unique category of the page. Emissions are focused on a multinomial distribution over word occurrences, as in the Naive Bayes classifier's generative portion. The HMM is trained from (partially) labeled page sequences, in essence, state variables in the training set are partially observed(Kang, Ahn, & Lee, 2017).

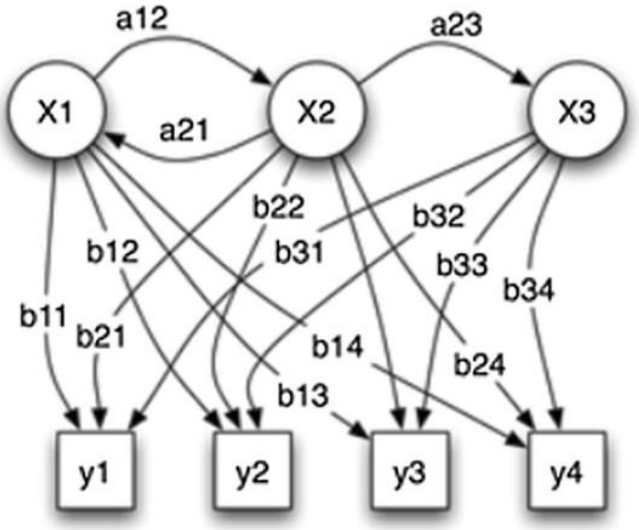



**Figure 2.5: Architecture of a hidden Markov model**

### 2.5.8.2 Discriminative

Discriminative methods used for uncontrolled and supervised learning as well as other specific algorithms are designed to be illustrative rather than comprehensive. It has often been argued that discriminative classifiers often achieve higher test set accuracy for many application domains. Support Vector Machine (SVM), Logistic regression, traditional neural networks, Nearest neighbor, conditional random fields (CRF) are popular models (Yogatama et al., 2017).

The nearest neighbor (NN) approaching data classification estimates the possibility of a data point being a member of one group or another depending on which group of data points are closest to it. The NN rule identifies the unknown data point category based on its closest neighbor whose class is already known. This rule is commonly used in applications for pattern recognition, categorization of text, ranking models, object recognition, and event recognition (Bhatia & Vandana, 2010). A NN classification scheme for text categorization using mutual information and weight adjustment techniques to learn the importance of discriminating words. It is for a particular document are then computed based on the matching words and their weights (Wan, Lee, Rajkumar, & Isa, 2012).

Support Vector Machine (SVM) is another powerful algorithm based on the theory of Vapnik-Chervonenkis, as defined by Oracle docs. This supervised machine learning algorithm is well regulated and can be used for classification or regression challenges. The support vector machine (SVM) was one of the best-known classification techniques and is usually the best classification for computer-assisted radiology detection (Larroza et al., 2015). Support vector machines have specialist properties that minimize the error of empirical classification and maximize the geometric margin at the same time. SVM map input vector to a higher dimensional space where a maximum hyperplane is built. The two parallel hyperplanes establish a distance between two parallel hyperplanes on each side of the hyperplane to isolate the information and divide the hyperplane. The assumption is made up of a greater margin or distance between the parallel hyperplane to generalize classifier error better (V. Vapnik, 1995). Figure 2-5 shows the total hyperplanes margin for an SVM equipped with two-class samples.

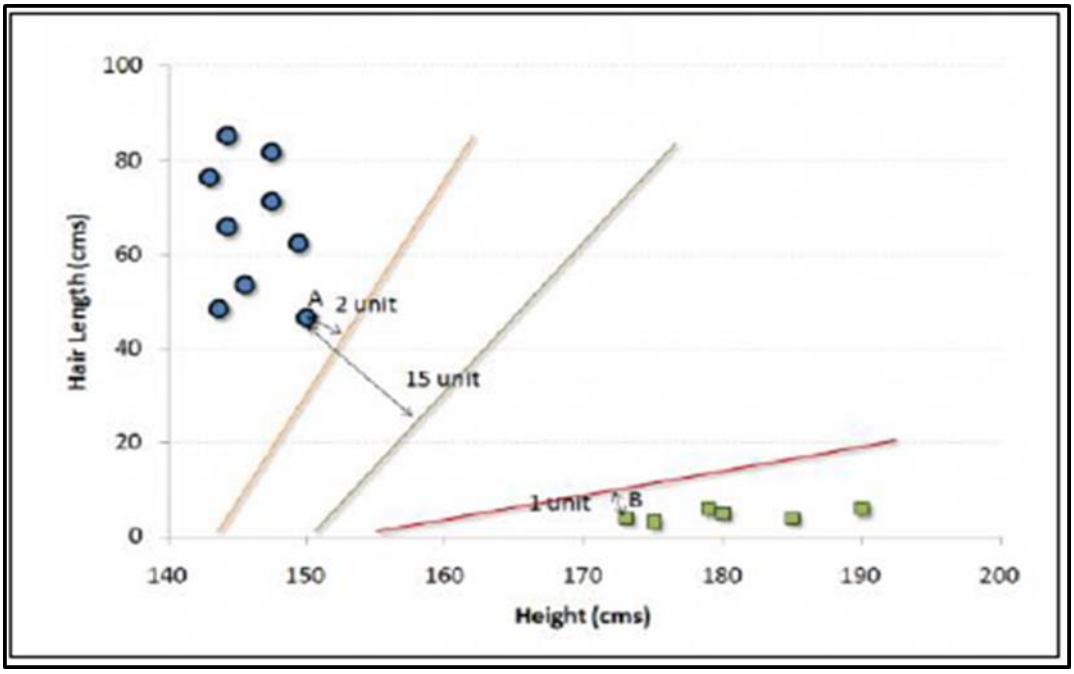



**Figure 2.6: Maximum margin hyperplanes for an SVM**

## 2.6 Critical Review

### 2.6.1 Previous Research on Spam

Online review forums are attracting significant attention from both individuals and businesses with the rapid growth of social network platforms and internet media. Because of its utility and impact on e-commerce markets, user-generated content is becoming increasingly valuable to both individuals and businesses. Based on research by Kim, Chang, Lee, Yu, & Kang (2015), propose a frame-based method of deep semantic analysis to understand the rich characteristics of misleading and truthful opinions written by different types of people. The researchers classified spam review into extreme forms (untruthful opinions, brand-only reviews, and non-reviews) by characteristics. The simple linguistic feature-based classification system (such as n-gram) is highly accurate in the detection of crowd-based spam opinion data sets. A linear kernel is used for the separation of the SVM hyperplane.

A research by Wang, Zubiaga, Liakata, & Procter (2015), focus on spam tweet detection, which optimizes the amount of data to be collected by relying on the inherent features of the tweet. This allows the spam detection system to be applied in a timely fashion to a large set of tweets, potentially applicable in the real-time or near-real-time setting. They used the Twitter spam datasets, the pre-processing techniques were used on the tweets and the four separate feature sets were used to train our spam vs. non-spam classification. Preprocessing techniques such as spell checking and stemming were tested but subsequently discarded due to the minimal effect and observed in classifier performance. To extract sentiment-based features for the specific case, remove hashtags, links and user mentions from tweet contents as well. N-gram models have been used for various tasks in natural language processing, including text classification. Powerful with a reasonable amount of training data for simple text classification.

A wide range of users finds the growth of social media, email services and other internet facilities helpful. In this, spam emails are one of the social network's most alarming behavior. Efficient spam filters are required in this case, and most email service providers dispense their own filtering mechanisms to manage spam mails. A study by Mathew & Ramani Bai (2017), a set of features created using N-grams

technique was used as a training set for Naive Bayesian classifier and its classification efficiency was analyzed for different gram ranges. After applying all the processing steps used in the Simple Naive Bayes Classifier, the N-grams generator co-operated in the pre-processing stage. Trigrams, 4-grams and 5-grams were evaluated for the classifier and the filter efficiency was analyzed for each event. In hybrid filtering techniques, the Naive Bayesian classification method can be used as a primary filter.

According to research by Reddy & Reddy (2019), social networking sites provide users with a great deal of technical information. This large amount of information on social networking sites is attracting cybercriminals to misuse the information on these sites. These users create their own accounts and distribute vulnerable information to the actual users. The researcher proposed a hybrid approach that uses content-based and user-based features to identify spam on Twitter network. They used a decision tree induction algorithm and Bayesian network algorithm to construct a classification system in this hybrid approach. Analyzed the proposed twitter data set technique. A proposed integrated solution was developed to improve the detection of spam messages, incorporating the advantages of the two classification algorithms. The spam detection improvement is calculated on the basis of the parameters of accuracy.

A research by Roy & Viswanatham (2016), spam emails have become a rising threat for all web users. Such unsolicited messages frequently drain network resources. This paper looks at the capabilities of the Extreme Learning Machine (ELM) and Support Vector Machine (SVM) to classify spam emails with class level. The ELM method is an efficient model based on a single-layer neural feed-forward network that can randomly pick weights from hidden layers. The SVM is a powerful theory of statistical learning, often used for classification. The comparative study examines precision, accuracy, recall, false positive, true positive and sensitivity analysis of ELM and SVM for the classification of spam emails. The findings identify the efficiency and effectiveness of the ELM and SVM models.

**Table 2.1: SPAM literature**

| Author | Tittle | Result/Description | Dataset |
|---|---|---|---|
| Seongsoon Kim, Hyeokyoon Chang, Seongwoon Lee, Minhwan Yu, Jaewoo Kang | Deep Semantic Frame-based Deceptive Opinion Spam Analysis | The Yelp data classification tests using the n-gram function yielded a 0.625 accuracy. This result is consistent with the fact that simple linguistic features are hardly effective for real-life datasets (at their Yelp dataset, an accuracy of 0.676 was obtained). Their experimental results show that for the AMT dataset, the classification model using frame features outperformed the baseline model by 4.34 percent. | Amazon Mechanical Turk (AMT) and Yelp dataset |
| Bo Wang, Arkaitz Zubiaga, Maria Liakata, Rob Procter | Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter | Results show that by using the limited set of features readily available in a tweet, they can achieve competitive results compared to existing spammer detection systems using additional, expensive user features. | Twitter spam datasets |
| Nikhil V Mathew, Ramani Bai V | Analyzing the Effectiveness of N-gram Technique Based Feature Set in a Naive Bayesian Spam Filter | The accuracy of the Naive Bayesian spam filter can be improved by using the n-grams feature set. Because it has a comparatively less false positive rate and optimal accuracy, the use of a 4-grams feature set is more efficient among the n-grams. | ENRON dataset |

| | | | |
|---|---|---|---|
| K. Subba Reddy, E. Srinivasa Reddy | Integrated approach to detect spam in social media networks using hybrid features | Used decision tree classifier induction, Naive Bayes classifier and hybrid features in this integrated approach, such as combining user-based features and content-based features. Their approach shows high performance, accuracy and F-measurement. | Twitter database |
| Sanjiban Sekhar Roy, V Madhu Viswanatham | Classifying Spam Emails Using Artificial Intelligent Techniques | ELM and SVM's performance is inspiring as both have achieved highly accurate detection rates in spam email. ELM and SVM can be used to solve various computer science classification problems | spambase obtained from UCI machine learning repository |

### 2.6.2 Previous Research on YouTube Spam

Recently, various unwanted threats have affected online social networks. The malicious users often post links, advertisements and fraudulent information on the phishing website in the comment section that can transmit viruses or malware. According to (Sharmin & Zaman, 2018), analyzed the latest concept and methods for spam comment detection posted on YouTube, one of the most popular social media sites. The main purpose of this study is to classify comments as spam or legitimate and to measure the performance of an ensemble classifier over a single classifier algorithm in the context of text classification. The term Frequency-Inverse Document Frequency (TF-IDF) was measured and given the relevance weight of the TF-IDF measures. Stop words removing and stemming were also used to improve the classification algorithm accuracy measurement. Classification is a supervised learning method that classifies class values. K-Nearest Neighbor, Naive Bayes, Support Vector Machine (SVM) and Bagging have been implemented. To verify the performance of the model, all

experiments are conducted with 10-fold cross validation. Using WEKA machine learning tool, this experiment was carried out. WEKA is an open source software that consists of data mining learning algorithms.

YouTube is one of the largest sites on the Internet for users to get information. That's why many spammers will trick the YouTube user by spamming the comments on YouTube. According to (Aziz et al., 2018), this study uses the Support Vector Machine (SVM) and K- Nearest Neighbor (k-NN) to develop a YouTube detection system. Five (5) stages, such as data collection, pre-processing, feature selection, identification, and detection, are involved in this study. This method is chosen as it can provide good accuracy for the test. This method also allows the comparison of the effects of the SVM technique and the k-NN technique. Weka and RapidMiner are used to perform the experiments. The dataset contained five videos selected and was downloaded through the API from YouTube. Data cleaning such as tokenization, removal of stopwords and stemming will be performed by the raw data set. The result shows that the performance of SVM in weka is better than finding in RapidMiner. The output for SVM is slightly different from k-NN which is 91.49% to 90.59% in Weka. While result in RapidMiner show that SVM is higher (74.40%) than k-NN (56.70%). These result can be seen on figure 2.6.

| | Weka | Rapidminer |
|---|---|---|
| **classifier** | **accuracy** | **accuracy** |
| k-NN | 90.59% | 56.70% |
| Support vector Machine | 91.49% | 74.40% |

**Figure 2.7: Results based two techniques used**

A research by Aiyar & Shetty(2018), proposed a new approach to identify unwanted comments or spam on the YouTube video sharing platform. The existence of these comments greatly affects a channel's credibility as well as the normal users ' experience. YouTube itself has very limited methods to tackle this problem by blocking comments that contain links. In this work, the researcher attempts to detect such comments by applying conventional machine learning algorithms such as Random Forest, Support Vector Machine, Naive Bayes and certain custom heuristics

such as N-Grams which have proven to be very effective in detecting and subsequently combating spam comments. In this direction, the aim is to attempt to classify algorithms and to apply heuristics such as N-Grams that can accurately detect spam with a high F1 rating and thus increase classification accuracy. Cross-validation and k-fold approach performance of spam comment detection system. Using a random number, the data set is shuffled. A five-fold process of cross-validation was used. Character-grams can better identify the spamming measurement in the comment as opposed to word-grams. YouTube application and the effectiveness of using word-grams (n character) over word-grams to improve classification accuracy. Support Vector Machines & Random Forests actually outperform other traditional classification machine learning algorithms and are very suitable for high-dimensional data set.

Next study by Alper Kursat Uysa (2018), Spam filtering is one of the most popular classification domains for text. To analyze the performance of five states of the art text feature selection methods for spam filtering on YouTube using two excellently-known classifiers, namely naïve Bayes (NB) and decision tree (DT). In the experiments, five datasets were used, including spam comments from different subjects. The measure of success of Macro-F1 has been used. Furthermore, for fair performance evaluation, 3-fold cross-validation is recommended. Filter-based selection methods are widely preferred for text classification because there are many features and these types of methods do not interact with classifiers during the selection process. The study uses five well-known filter-based text selection approaches. This included the information gain (IG), the Gini index (GI), the distinguishing feature selector (DFS), discriminative features selection(DFSS) and the relative discriminative criterion (RDC).

**Table 2.2: YouTube SPAM literature**

| Author | Tittle | Result/Description | Dataset |
|---|---|---|---|
| Zakia Zaman, Sadia Sharmin | Spam Detection in Social Media Employing Machine Learning Tool for Text Mining | It was found that the accuracy level for most cases is above 80%, except for the Katy Perry dataset using the SVM algorithm, which is below 60%. In most cases, Naive Bayes and Bagging classifiers provide high precision measurements in this experiment. K-Nearest Neighbor ( 1 neighbor) also gives high precision. Nevertheless, reliability can be viewed as a great performance indicator when having symmetrical data sets where the values of false positives and false negatives are almost equal. | YouTube Spam datasets from UCI data repository |
| Aqliima Aziz, Cik Feresa Mohd Foozy, Palaniappan Shamala, Zurinah Suradi | YouTube Spam Comment Detection Using Support Vector Machine and K–Nearest Neighbor | The output for SVM is slightly different from k-NN which is 91.49% to 90.59% in Weka. While result in RapidMiner show that SVM is higher (74.40%) than k-NN (56.70%). It is recommended that to use Weka detect spam comments on YouTube. Weka provides more accuracy. | YouTube Spam datasets from UCI data repository |
| Shreyas Aiyara, Nisha P Shetty | N-Gram Assisted Youtube Spam Comment Detection | Character-grams can better identify the spamming measurement in the comment as opposed to word-grams. YouTube application and the effectiveness of using word-grams (n character) over word-grams to improve classification accuracy | Open Youtube API from different Youtube channels |

| | | | |
|---|---|---|---|
| Alper Kursat Uysal | Feature Selection for Comment Spam Filtering on YouTube | Their studies have shown that most of the highest classification quality has been obtained with DFS and GI feature selection methods. Nonetheless, for some cases, DFSS and RDC tend to be less successful than the others. | YouTube Spam datasets from UCI data repository |

## 2.7 Conclusion

In a nutshell, this chapter provides a description of the area that this work will use and focus on. It is can be done with study the past literature review that produce by previous researcher. It seen in order to focusing in YouTube spam detection, the features selection should be done with using Information Gain to reduce the number of features that have redundancy. After that, classification with support vector machine will also produce results based on the accuracy. The datasets used for this research are obtained from the UCI Machine Learning Repository. Solutions are proposed for this research based on the fields discussed in this chapter. In the next chapter, the proposed methodology will be explained further.

# CHAPTER 3: PROJECT METHODOLOGY

## 3.1 Introduction

In this chapter, will explain in detail about methodology and approach that will used in this research. Besides that, This project will create a framework based on various key areas addressed in the previous chapter. The framework is designed to ensure that the research is conducted and carried out by following the appropriate steps and procedures. It also attempts to explain the method and flow chosen to complete the study. However, these phases of this research project will be explained in the section below. Finally, flowchart, milestone and Gantt Chart of this research project will also be included every progress in completing this research that can be accomplished according to the given timeline.

## 3.2 Methodology

Methodology of the project that is important to ensure that this project is carried out on the basis of the proposed flow. The methodology is used in a theoretical analysis of processes that serve as a basis for explaining how the system can accomplish the objectives of the project. In this research, framework that are chosen according best machine learning in YouTube spam detection. The use of machine learning to predict on a dataset that will subsequently be classified for the accuracy or false positive results. Otherwise, there are six phases to be implemented in this research and to ensure the smooth operation of the project. The six phases involved are previous research, information gathering, define scope, design and implementation, testing and evaluation of system, and documentation. The methodology of the system is represented in the illustration shown below. Figure 3-1 shows the framework of the methodology model involved in this project.

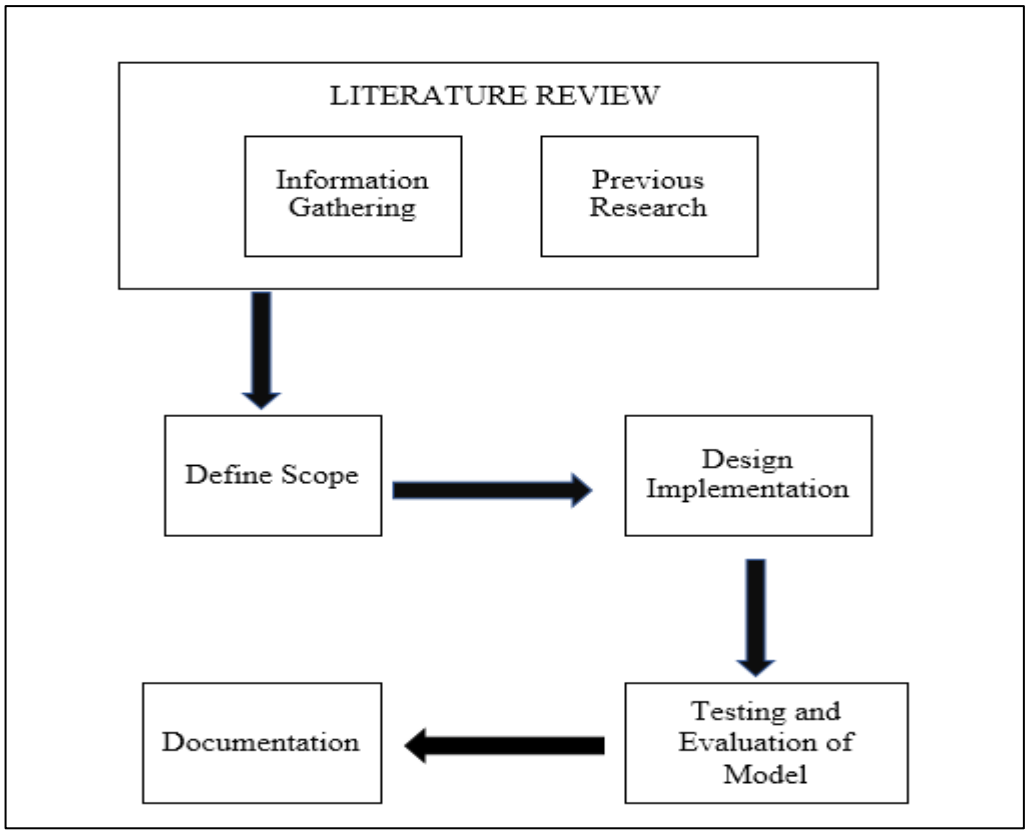



**Figure 3.1: Framework of the System**

### 3.2.1 Previous Research

This phase provides a better understanding of how to conduct the project on the basis of previous research. It is important to ensure that previous research has defined all requirements. Because the domain is required before the project is carried out to address the defined projects. There are also many domains to be known as spam types, learning machines, features extraction, features selection and classification method. Previous research will provide insight into how the planned theoretical frameworks operate in their respective fields. The next step will describe the architecture, framework and related domain architecture in more detailed information.

### 3.2.2 Information Gathering

Information gathering aims to have a solid understanding of research problems. Once all domains have been identified, this phase will be taken to determine the issue of internet security attack in question. The detection system and the types of features used will also be listed. Spam means the disruption and be within the sort of advertising or similar specific content which can contain malicious code hidden in it. In addition, the YouTube may have legitimate or spam content where spammer used this chance to spread malware across comment fields. On the other hand, this study will develop a multi-class classification system using Machine Learning. The gathering of information from previous research will improve the choice of algorithms and methods used in this research to conduct experiments.

### 3.2.3 Define Scope

The scope of the project is the limit of where the research is carried out. This research conducted to analysis the data of YouTube spam. Therefore, this phase will focus on reviewing the YouTube spam detection by using a content that is contained in the comment. The dataset a source from UCI machine learning repository which focus on dataset. Subsequently, the study will focus on the types of YouTube spam attack that can be found in the current era. Meanwhile, new methods are proposed to produce accuracy or false positive results in detecting spam comments.

### 3.2.4 Design and Implementation

This phase will describe the design and implementation of the proposed new models based on the previous phase study. This focuses on the design of the new method of detection. Firstly, the dataset is gained from the UCI Machine Learning Repository. Then, the data set is separated into a few considerate features for this research. Next, the data set is extracted using N-gram features to be trained. The new feature set is then used in classification using SVM in which developed the newer model of detection.

### 3.2.5 Testing and Evaluation of Model

The investigation will be carried out in detail during this test phase so that the system can be defined. This is because the system does not function properly, process failures are identified in the system. This test phase also plays an important role in ensuring that systems meet the user's needs and function effectively. After evaluation, the lack of a system that is informed by the user will be recorded and the new system improved periodically. This new system is develop to meet the needs of consumers. The consideration also will be given to the new system, occurrences of this phase will be made to meet the system's lack.

### 3.2.6 Documentation

The documentation process helps to organize the results in a more organized and proper way. Documentation is important to help researchers improve and enhance this research project in order to provide research ideas. Each experiment is properly written, restricts all procedures and results in proper documentation that serves as a reference and proof of each activity. It's supposed to be a headache without proper documentation strategy to keep track and manage it all. In each respective sections, all relevant information is first identified and documented.

## 3.3 Project Schedule and Milestones

### 3.3.1 Project Flowchart

A flowchart is used for a summary of the tasks and the relation in a systematic way. This process will also define the necessary resources to be mapped with their respective activities in order to prevent delays or potential constraints. Figure 3-2 shows the flowchart of the general phases that are involved in this project.

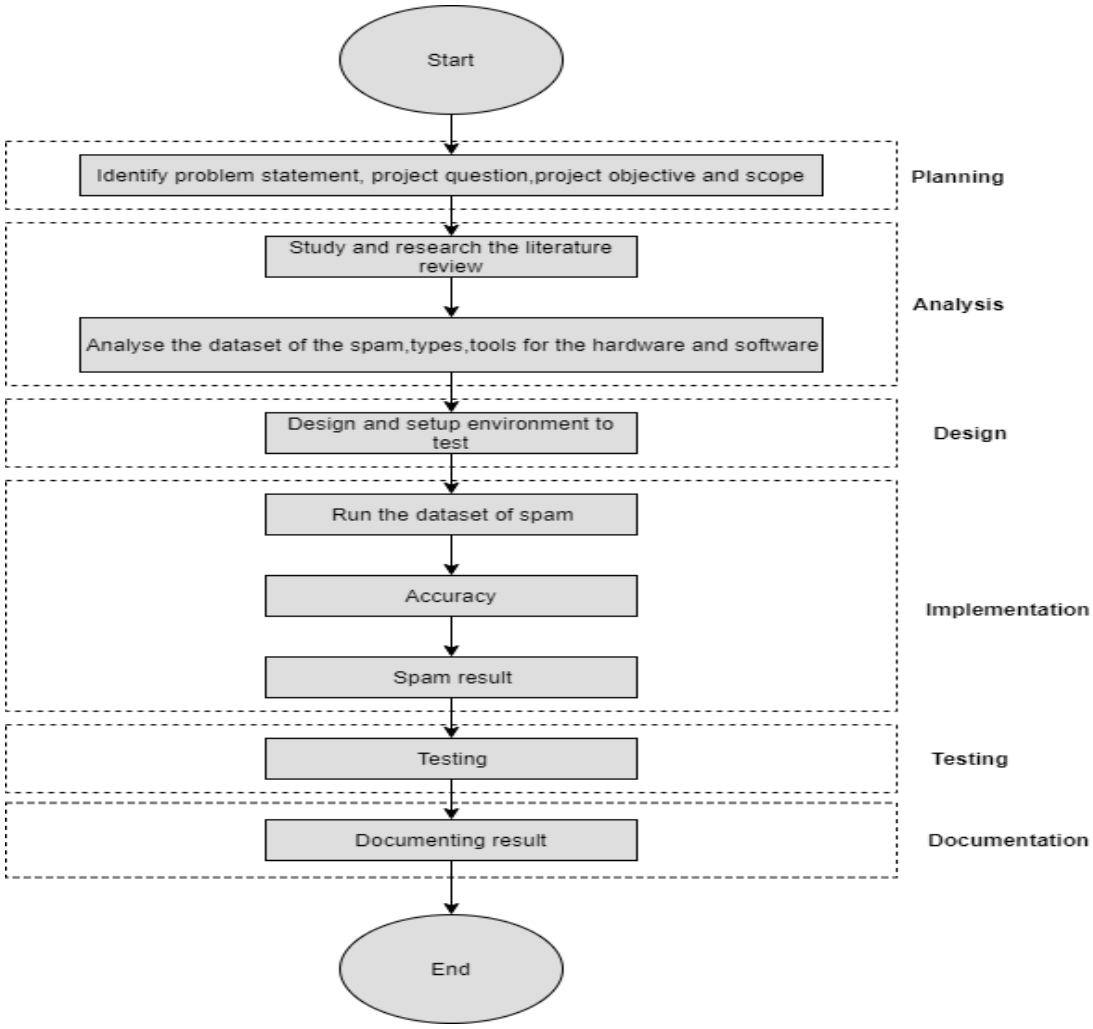



**Figure 3.2: Project Flowchart**

### 3.3.2 Project Milestones

The project milestone is important for monitoring upcoming events or targets throughout the specific timeline. Significant value could be applied when using milestones as it allows us to determine whether or not the plan is progressing on the basis of the schedule. Table 3-1 below shows the project milestone for this research.

**Table 3.1 Project Milestone**

| Week | Phase | Action | Deliverables |
|---|---|---|---|
| 1-5 | Planning | (9/9/2019 – 13/9/2019)<br><br>Identify title, problem statement and scope. | Complete Proposal |
| | | (16/9/2019) – (20/9/2019)<br><br>Study and research the literature review. Write and submit project proposal to supervisor. | |
| | | (23/9/2019) – (27/9/2019)<br><br>Proposal accepted. | |
| | | (30/9/2019) – (4/10/2019)<br><br>Identify title, problem statement, objective and scope of project | Chapter 1: Introduction |
| | | (7/10/2019) – (11/10/2019)<br>Chapter 1 is done and submit to supervisor for evaluation. | Progress report<br>Chapter 1 |
| 6-9 | Analysis | (14/10/2019) – (18/10/2019)<br>Studies on related work and previous research and finding taxonomy of spam classification. | Chapter 2: Literature Review |
| | | (21/10/2019) – (25/10/2019)<br>Study methodology on previous research. | Chapter3: Methodology |

| | | | |
|---|---|---|---|
| | | (28/10/2019) – (1/11/2019) | MID SEMESTER BREAK |
| | | (4/11/2019) – (8/11/2019)<br>Information collection and analysis. | Chapter 4: Analysis and Design |
| 10-15 | Design | (11/11/2019) – (15/11/2019)<br>Design the project and choose the tools for implement | Chapter 4: Analysis and Design |
| | | (18/11/2019) – (22/11/2019)<br>Design the environment for implementation. | Progress report on Chapter 4 |
| | | (25/11/2019) – (29/11/2019)<br>Write and finalize project report. | PSM1 Report |
| | | (2/12/2019) – (6/12/2019)<br>Submit project report to supervisor. | PSM1 Report |
| | | (9/12/2019) – (13/12/2019)<br>Schedule the Presentation | Presentation Schedule |

### 3.3.3 Project Gantt Chart

Gantt chart is a timeline for each activity that can be found throughout the project. When an activity included in the project is late, it will affect the future durability and the costs expected to increase over the remainder of the activities involved in the project.

Refer to appendix section.

## 3.4 Conclusion

In conclusion, The methodology is the most important stage in the development of the project. If a project does not have a methodology the project faces challenges when it is implemented. As a result, all steps involved in this chapter will be discussed so that during the research conducted, the approach taken to make observation can be made. The accuracy or false positive for detecting potential spam comments on YouTube is measured using the methodology suggested in this chapter. The next chapter will discuss in detail the method process and steps of the design a N-gram features of learning machines using SVM.

# CHAPTER 4: ANALYSIS AND DESIGN

## 4.1 Introduction

In this chapter, the detailed explanation of the design and analysis of the methods to be used or implemented in the project is running. To develop the system this phase requires full concentration. In addition, one continuation made upon recommendation of the methodology were discussed in the previous chapter. The analysis will be carried out on the project to identify the system's interests and needs with the purpose of the system being created. The method of chi-square and information gain is identified as the best method to be used in this experiment. The two techniques were selected based on the literature's proven excellent results. The requirement for software and hardware is also evaluated in order to determine the best system for this study. The design phase includes further analysis of each of the methods used for the research.

## 4.2 Problem Analysis

The main objective of this research is to build a model of machine learning to differentiate between spam and legitimate comment on YouTube. The process of detection is done automatically using the pre-program algorithm of machine learning. The problem is when the data is not in a binary structure that is legitimate and infected and there are five features in the data set that are comment id, author-date, content, and class. This research will find features that will be used and the content and class have been chosen due to other features that are difficult to distinguish between spam and legitimate comment to accommodate all five features. Many instances require high computing resources and a high-precision algorithm for machine learning. This research will solve the classification of YouTube spam . The classifier method used is the SVM, which was one of the best-known classification techniques and is usually

the best classification for computer-assisted detection. This technique is the best way to solve problems

## 4.3 Project Design

The design of this project will show that there are steps to be carried out in the research projects. Therefore, each phase has a distinctive function in processing the information in order to make it a decision according to the goals of this research project. Illustration Figure 4.1 phase project generally shows the concept of machine learning standard method for detection program of comment spam. The process begins with data pre-processing, extraction, normalization and classification. First of all, this requires that analysts have been identified using a machine learning project research dataset. Data set is the raw data collected as a special research topic and used to convey information to the machine learning algorithms to detect spam comments.

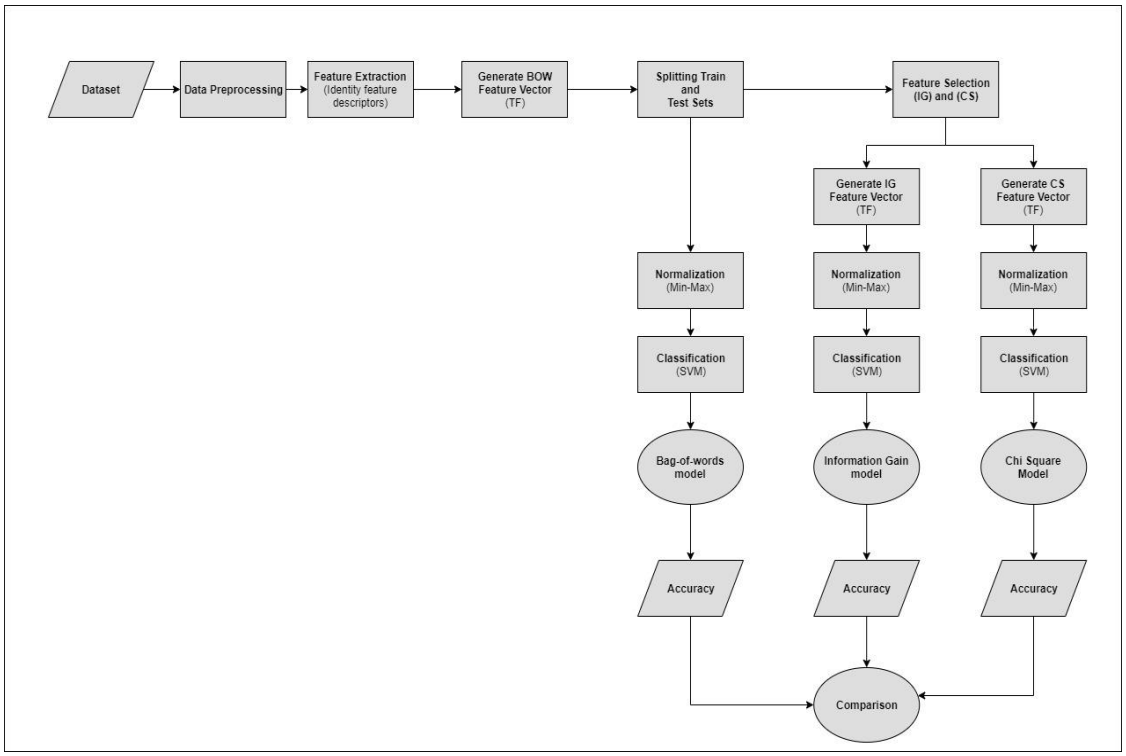



**Figure 4.1: Project Design**

### 4.3.1 Dataset

A data set is required to evaluate and apply the concept and theory of the stated machine learning algorithm. The dataset is the raw data obtained in related to this research topic. It is used to provide the information to the machine learning algorithm to detect spam comments. This dataset is commonly used by the other researcher of YouTube spam. The datasets used in this paper obtained from the Machine Learning Repository's YouTube Spam Collection Data Set. This data set was donated on 26 March 2017 and the data collected using the YouTube Data API v3. Five YouTube video comments were obtained in the datasets with a total of 1956 comments. A total of 1005 comments are spam, while the rest are legitimate comments. There are five features in the data set that are comment id, author-date, content, and class as shown in the figure 4.2. Figure 4.3 shows features that will be used as the content and class.

**Table 4.1 Description of Dataset Shakira**

| Dataset | Number of Spam | Number of Legitimate | Total comment |
|---|---|---|---|
| Psy | 175 | 175 | 350 |
| Katy Perry | 175 | 175 | 350 |
| Eminem | 245 | 203 | 448 |
| LMFAO | 236 | 202 | 438 |
| Shakira | 174 | 196 | 370 |
| TOTAL | 1005 | 951 | 1956 |

| COMMENT_ID | AUTHOR | DATE | CONTENT | CLASS |
|---|---|---|---|---|
| z12lhjtxlsisx55y | Shirley Lim | | Take a look at this video on YouTube:ï»¿ | 1 |
| z13hhxajgrnldjn | TMCB production (Insti | | Check out our Channel for nice Beats!!ï»¿ | 1 |
| z12xurdj0qznep | Sylith | 2015-05-2 | Rihanna and Eminem together are unstoppable.ï»¿ | 0 |
| z13isdtoikvzjlkij | nepaladventure | | Check out this playlist on YouTube:ï»¿ | 1 |

**Figure 4.2: Example of Eminem dataset**

| CONTENT | CLASS |
|---|---|
| Take a look at this video on YouTube:ï»¿ | 1 |
| Check out our Channel for nice Beats!!ï»¿ | 1 |
| Rihanna and Eminem together are unstoppable.ï»¿ | 0 |
| Check out this playlist on YouTube:ï»¿ | 1 |

**Figure 4.3: Example of features will be use**

### 4.3.2 Data Preprocessing

Pre-process refers to activities that transform raw data into data that the machine learning algorithm easily understands. Pre-processing is the first step in which suspicious sources and documents are subject to specific improvements such as the removal of the stop word, tokenization, lowercasing, stemming, removal of special character, etc. According to the Vani & Gupta (2014),this will helps to reduce the size of actual data by removing irrelevant information about the method used. In this research the attribute selection is comment and the process will focus on the content.

The method of tokenization is the method of separating the raw text into sentence and then into tokens. Most of the words in documents repeat very often but are basically meaningless as they are used to combine words in a sentence. Word tokenization for numeric data transformation becomes a critical part of the text (string). The tokens can be words or numbers or points for punctuation, etc. Split the texts into words or smaller subtexts so that the relationship between the texts and the labels can be generalized well.

| Preprocess CONTENT |
|---|
| take a look at this video on YouTube Check out our Channel for nice Beats ! ! Rihanna and Eminem together are unstoppable . Check out this playlist on YouTube : |

**Figure 4.4: Tokenization**

Stop words are often occurring and insignificant words in a language that helps build sentences but does not represent any document content. The common English stop words include a, about, an, are, at, be, by, for, how, etc. Stop words commonly not contribute to the context or content of textual documents. These stop words are not useful for document classification because it has in spam and legitimate class.

| Preprocess CONTENT |
| --- |
| take look video YouTube Check Channel nice Beats ! ! Rihanna Eminem unstoppable . Check playlist YouTube : |

**Figure 4.5: Remove stop words**

The data set collection may contain the special character and has both lower and upper cases. These special character and lower case had to be preprocessed and properly organized. This included removing full stops, semicolons, symbols, quotation marks, and converting all words into lower cases.

| Preprocess CONTENT |
| --- |
| take look video YouTube Check Channel nice Beats Rihanna Eminem unstoppable  Check playlist YouTube |

**Figure 4.6: Remove special character**

| Preprocess CONTENT |
| --- |
| take look video youtube check channel nice beats rihanna eminem unstoppable  check playlist youtube |

**Figure 4.7: Case Normalization**

The stemming method is refers to the reduction of words to their stems or roots. Stemming allows for the retrieval of different variations of the word, which improves the recall. By deleting affixes, Stemmer reduces the word to its root . The purpose of stemming is to save time and space to efficiently and quickly make classification (Rana, Khalid, & Akbar, 2015). This preprocessing step will help in cases where the data set is not very large and significantly contributes to the reliability of the expected

output. Other than that it helps reduce the number of features that help maintain a decent size of the models.

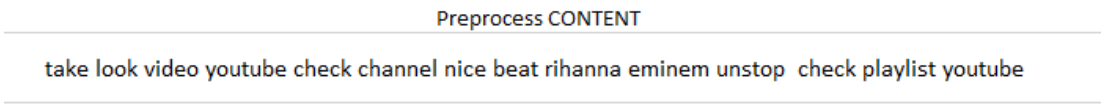



**Figure 4.8: Stemming**

After finish all the preprocessing rule step the feature descriptor is generate as show in figure 4.8. Table 4.2 indicates the advantage and disadvantage implementing the rules for this study.

**Table 4.2: Advantage and disadvantage of rules**

| Rules | Advantage | Disadvantage |
|---|---|---|
| Tokenization | Separating the raw text into sentence and then into tokens. | It will make it impossible to generalize a good relationship between the texts and the labels. |
| Stop Word Removal | Stop word removal enhanced classification accuracy in most cases. | The words take up space in the database or retrieve valuable processing time |
| Stemming | Transform words to their roots and to reduce the dimensionality of the feature space and improve the efficiency of the text classification system | Stemming may cause the same word to stem two different words. |
| Case Normalization | upper-case or lower-case word forms are considered to have no difference. fewer features are | Sentence with a capital letter separate from the same word that appears later in the sentence, but without any |

| | | |
|---|---|---|
| | accomplished with more discrimination | capital. This may result in a decline in accuracy. |

### 4.3.3 Feature Extraction

Feature extraction starts with the initial set of measured data and constructs derived values features intended to be informative and non-redundant. In this research will use n-gram to develop features. An n-gram is a contiguous set of n words from a single text string. In addition, the n-gram size that will be use is n=1,2,3,4,5. This method will identify the feature descriptor. Figure 4.9 shows the example of features descriptor using trigram which is n= 3.

```
tak ake loo ook vid ide deo you out utu tub ube che hec eck cha han ann nne nel nic ice bea eat rih iha nna emi min ine uns nst sto top pla lay ayl yli lis ist
```

**Figure 4.9: Feature descriptor using 3-gram**

### 4.3.4 Generate Bag of Word (BOW) feature vector

BOW is generated as the benchmark for comparing its accuracy with the new model developed at the end of the experiments from this research. BOW is known as the vector space model that generates BOW important for calculating the frequency characteristics in the dataset. Model BOW is used to preprocess the text by translating it into a BOW that contains a list of the total occurrences of the most commonly used words. Dataset will collect from the UCI Machine Learning Repository to produce BOW and divide the data set into 2 parts, which is the first part for the train and the second part for testing. This phase will be done by using the probability output from the previous phase to generate a new feature vector. For this research, the implement method is the term frequency (TF), it refers to the frequent number of words in a document. The easiest option is to use a raw count for a word in the document.

| tak | ake | loo | ook | vid | ide | deo | you | out | utu | tub | ube | che | hec | eck | cha | han | ann | nne | nel | nic | ice | bea | eat | rih | iha | nna | emi | min | ine | uns | nst | sto | top | pla | lay | ayl | yli | lis | ist |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

**Figure 4.10: Example BOW feature vector**

### 4.3.5 Splitting Train and Test Set

In this splitting train and test set for this research use the method random sub sampling. Random sub-sampling is used to split the data set into a user defined subset. The training set includes a defined value and the model learns about this data to generalize it to other data. The test dataset (or subset) to test the prediction of the model for this subset. According to the previous researchers, it has been shown that random subsampling results in test data predictions are consistent compared to cross-validation. For this purpose, the data of the YouTube Spam comment dataset is divided equally as shown in the figure below.

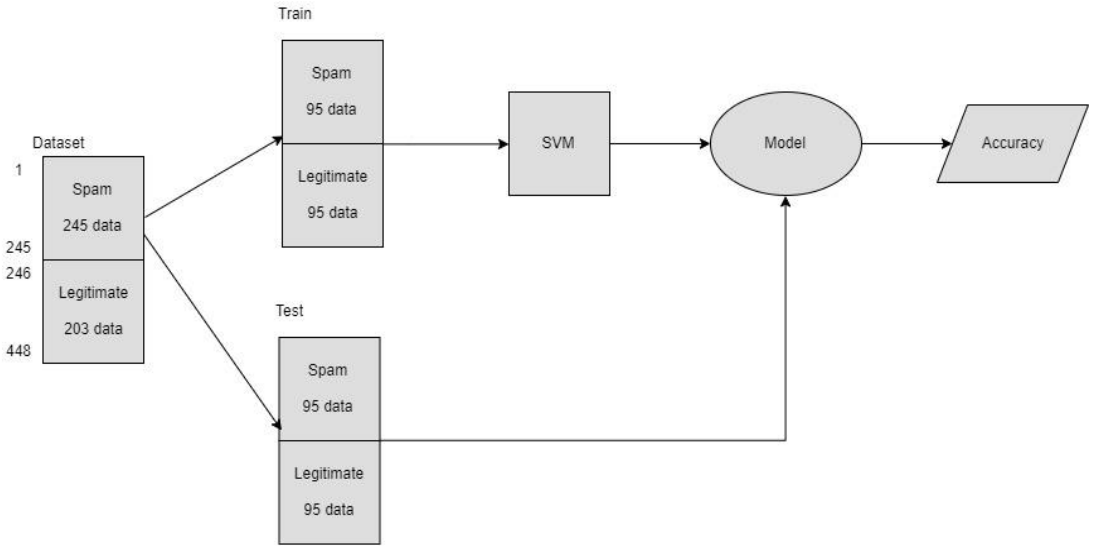



**Figure 4.11: BOW process using random subsampling**

Since the data set is finite, two methods can be used to process it. First, all the training data were used to select the best classifier and determine the error rate. This problem is related to two fundamental issues that are likely to over fit the training data, and the error rate is too optimistic or below the true error rate. In addition, the second solution for this project has been developed and used to overcome it. The method is to use random subsampling to divide the training data into disjoint subsets. A set number of instances is randomly selected for each class. The instances that has been chosen are not repeat which help to get a realistic estimation of the validity of the dataset.

As shown in figure 4.11 the total number of instances from both spam and legitimate data is 448. A data set will be randomized within a specific range to prevent randomization of different value classes. The random subsampling technique is then used to generate 190 random instances for train and test data. Then it represent the random instances that has been choose which is 95 instances for spam and another 95 instances for legitimate comment. The train and test data generated to train the classification algorithm fairly with the balance number of instances to detect spam and legitimate data, thus able to produce a fair result.

After the training process, it will determine the candidates from a set of parameters and also confirm the best achievements as known as the best kernel. The model will generate from the data model to be trained and generated using the SVM algorithm. Therefore, the model produced from the SVM method can provide accurate results and the time required to produce a model that can be reduced. When performing the test data, the data will be used as a form of model predictions. Therefore, the model to be used is the baseline forecast spam comment. The model form will be used as a prediction while the tested data will be used to detect spam comments.

### 4.3.6 Feature Selection

In this phase, will explain how CS and IG functions. CS and IG are evaluated during the feature selection process to verify their output to assess predictive accuracy. The attributes or features to be analyzed in the data train using CS and IG will be tested in order to achieve a performance that affects predicted accuracy. The given set of features dataset that is regularly spoken as a feature vector.

CS statistical works by performing the distribution class hypothesis test and independently evaluating all variables. CS attribute evaluates the value of a feature by calculating the value of the CS statistics for the class. Score is given based on the test done which follows the CS distribution . The test is carried out in order to rank the input features according to their priority. The initial hypothesis H0 is that the two features were unrelated and tested using the CS formula (Thaseen & Kumar, 2016). The number of main attributes required to predict the class variable is selected based on the value of the attribute. The following formula is used to obtain the value of CS.

$$\boldsymbol{X^2(t,c)} = \left[\frac{\boldsymbol{N \times (WZ - YX)^2}}{\boldsymbol{(W+X)(W+Z)(W+X)(Y+Z)}}\right]$$

Where:

W = No. of times feature t and class label c , X = No. of times t

Y = No. of times c , Z = No. of times neither c nor t

N = Total number of record

The IG feature selection is carried out in Weka. Weka is a collection of algorithms for machine learning for data mining tasks. This provides methods for data preparation classification regression, clustering, association rules mining, and visualization. Using the Ranker Search method, the Attribute Evaluator is used to calculate the information gain (entropy) for each output variable attribute. The entry values different between 0 (no information) and 1 (maximum information). The attributes that provide more data will probably have a higher rating for IG and will be chosen whereas one with less information will have a lower score and will be discarded. The following formula is used to obtain the value of IG.

Entropy:

$$H = -\sum_{i=1}^{k} P_k \log_2 P_k$$

Information Gain:

$$G = H - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$

Where:

k = Class H = Number of inputs P_k = Dataset m = total number of instances

IG is a symmetrical measure. In the wake of watching, the data increased about $m_L$, $m_R$ is equivalent to the increased data about $m_R$ after watching $m_L$. The shortcoming of the IG foundation is that it is one-sided when it is not more instructive for elements with more values than stand. For this research, the threshold of features selection will be set to 80% , then it will selected by using CS and IG. In the next phase, both features selected from two methods, CS and IG are used and compared which is best in this research to produce better predictive accuracy

### 4.3.7 Normalization

Normalization is the process of changing the data to a smaller range. In this research, the normalization is used to scale data from 0 to 1. For classification algorithms, normalization is mostly required. It changes the numeric column values in the dataset to use a more clear and understandable and typical scale without affecting the value range. There are many methods to normalize data such as min-max normalization. It is defined as shown in the formula.

$$y = \frac{(X - \min)}{(\max - \min)}$$

Where min and max are the minimum and maximum values of X and X is the set of the observed value.

### 4.3.8 Classification

Support vector machine or SVM that is one of the algorithms in supervised machine learning used in regression or classification. This SVM will be used in the classification phase for this project classifier as spam and legitimate comment. As in Figure 4.12 , SVM plots each data item as an n-dimensional space point (n = number of features) with a coordinate value for each feature.

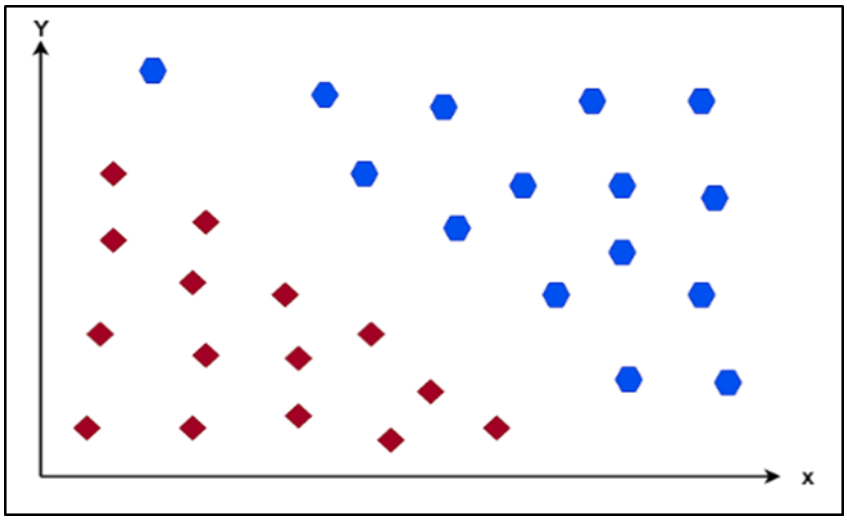



**Figure 4.12: Scatter plot**

In this research, the machine learning classifier using the SVM learning algorithm looks at how data can be classified and analyzed in a variety of ways. In addition, SVM uses hyperplane to plot the graph and model of the equation. The process of classification using SVM is shown in figure 4.13 below.

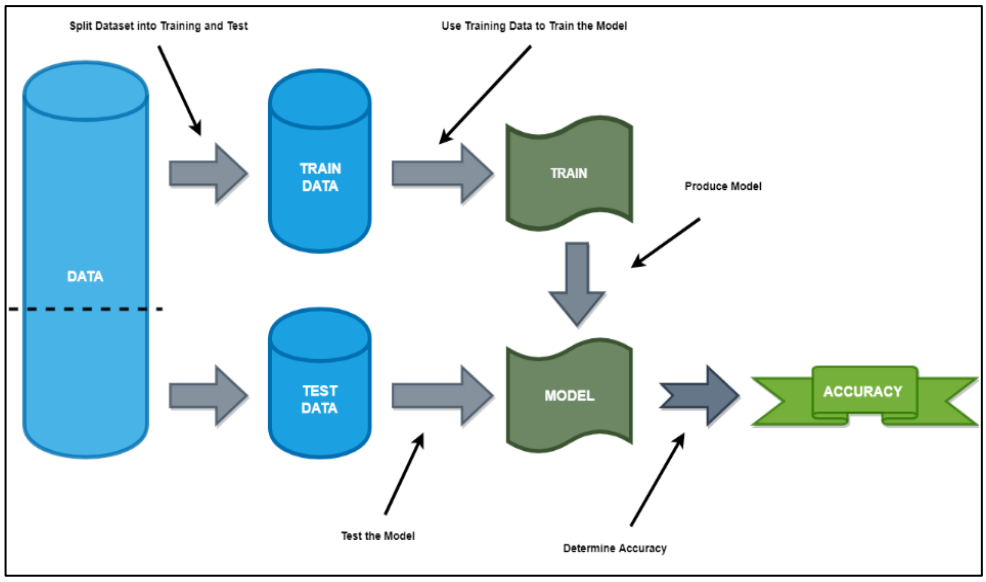



**Figure 4.13: Classification process using SVM**

After a plot has been identified, the SVM classifier will show how to plot in the graph. Plotting by using a classifier consisting of two parts, the first part being train data and the second part being test data. In separating the specific pattern sequence of class, the classifier will place the hyper plane. The classification using hyperplane shown in figure 4.14.

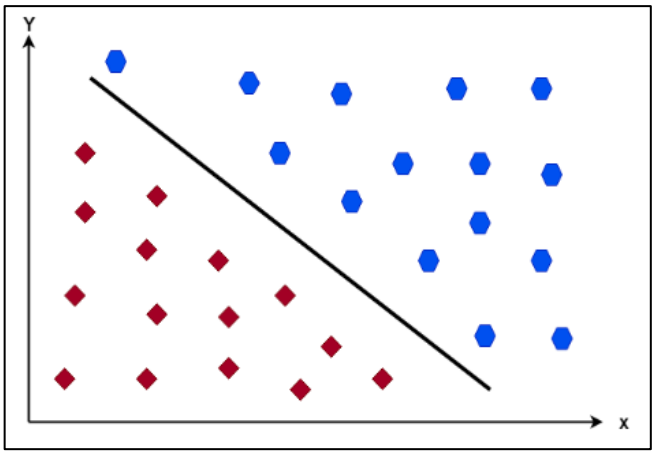



**Figure 4.14: Hyperplane Placement**

There are several functions can be implement in classification. One of the functions used in this research is the grid search task. Grid search function which searches for the best place of the hyperplane to distribute the vector area. In addition, grid search which is the process of scanning data to configure a parameter that give as

a model. Grid search can be applied to machine learning in calculating the best parameter to use a variety model. Grid-Search will create a model for each possible combination of parameters.

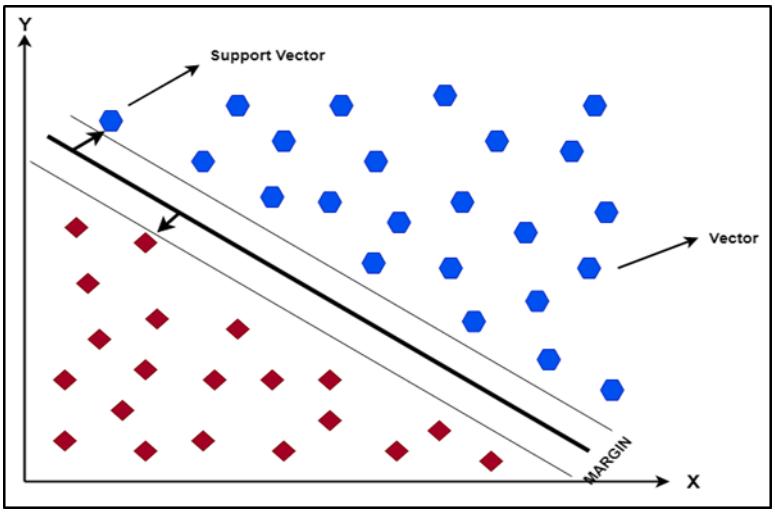



**Figure 4.15: Hyperplane with support vectors**

Figure 4.10 shows the hyperplane diagram with support vectors (at line margin and margin. In the SVM graph, show each plot in the graph that called in SVM as vector which is produce by train and test data. Following the completion of the plot, SVM will create a margin for finding the closest vector with margin and declared as a support vector. The problem that it is difficult to choose the best to get a hyperplane position.

For example, it seems that margin is used to try and find the right hyperplane by referring to the basic things in the SVM that choose the best margins. Figure 4.15 shows that hyperplane 2 is the classification of offenses and that hyperplane 1 is true. After that, the figure shows some predictable error. The hyperplane 1 accuracy higher than hyperplane 2 indicating a lower margin.

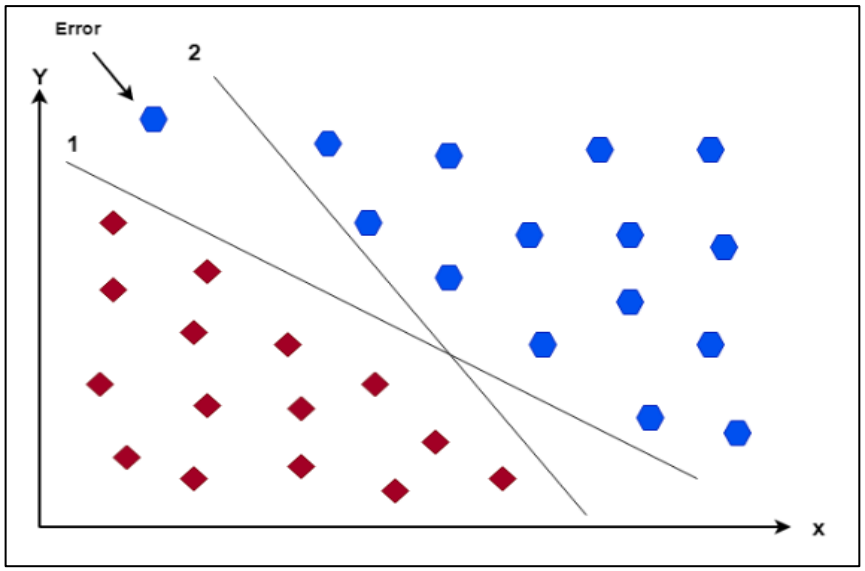



**Figure 4.16: Best hyperplane chosen**

It often occurs in the realm of human life that the error is conscious or unconscious. But the it has a solution to the error. Therefore, the data is often not predicted by the SVM itself, otherwise, the data generated will not make it logical. A confusing matrix method can solve the problem. This confusion method is a technique used to summarize a classification algorithm's performance. It also shows how confusing the model of classification is when making predictions..

This breakdown is the only way to overcome the limitation of using the accuracy of classification. This number of correct predictions and number of incorrect predictions are then organized into a table, or a matrix that is expected to follow down the side that is each row that matrix corresponds to a predicted class and another that is predicted over the top that is each column of the matrix corresponds to an actual class and then counts the correct and incorrect classification. Furthermore, the total number of incorrect predictions for a class goes into the expected row for that class value and for that class value the predicted column. Figure 4.17 below shows an example of real distribution

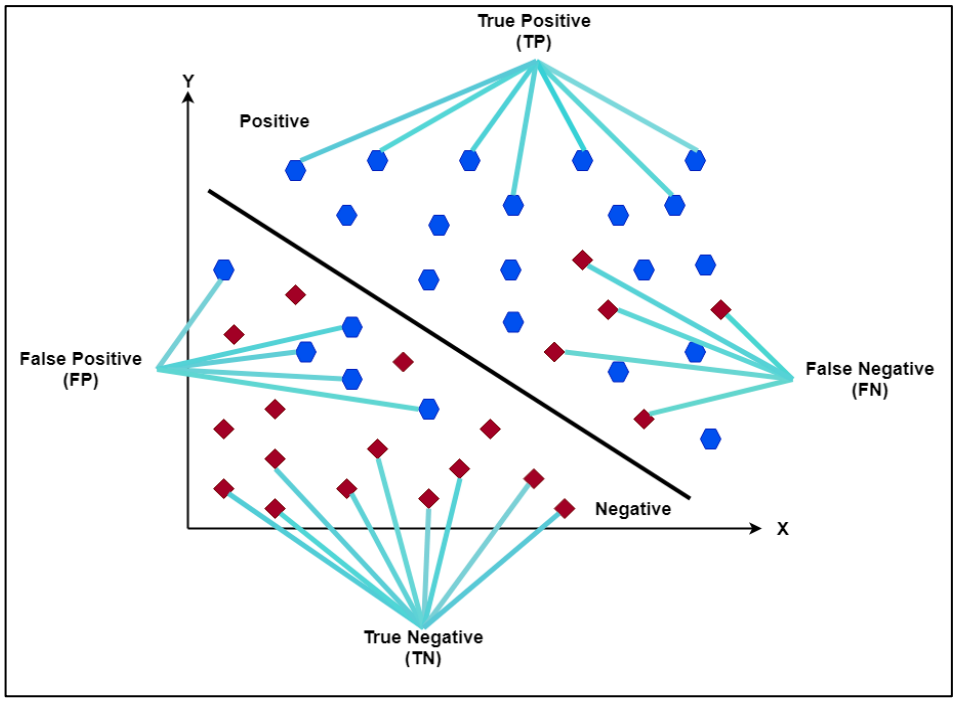



**Figure 4.17: Example of real distribution data**

Matrix of confusion is a combination of 4 different actual type and value prediction. This combination of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). True positive is that interpretation is predicted to be positive and identified to be true, while false positive is predicted to be positive, but it was false. For it has been predicted that true negative interpretation is negative and exactly false, while false negative interpretation is negative and true. In the security environment, fear becomes when false negative values are higher than false positive values. If this assessment is not carried out accordingly, it will affect security. These can describe the expected value as positive and negative as well as the true and false actual value. In confusion matrix can do several types of calculations for accuracy, accuracy, recall and f1 score. The formula that can be used as show below:

$$\textbf{Accuracy: } \frac{TP+TN}{Total\ Data} \qquad \textbf{Precision: } \frac{TP}{TP+FP}$$

$$\textbf{Recall: } \frac{TP}{TP+FN} \qquad \textbf{F1 Score: } 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

In figure 4.13, this case showed that one of the data in other areas was called an outlier. To solve this problem, the SVM has to ignore the maximum margin obtained and find the best hyperplane. The case SVM has the advantage of emphasizing the problem in solving it. The scenarios shown in the above description are Liner SVM dividing the vector by using straight lines.

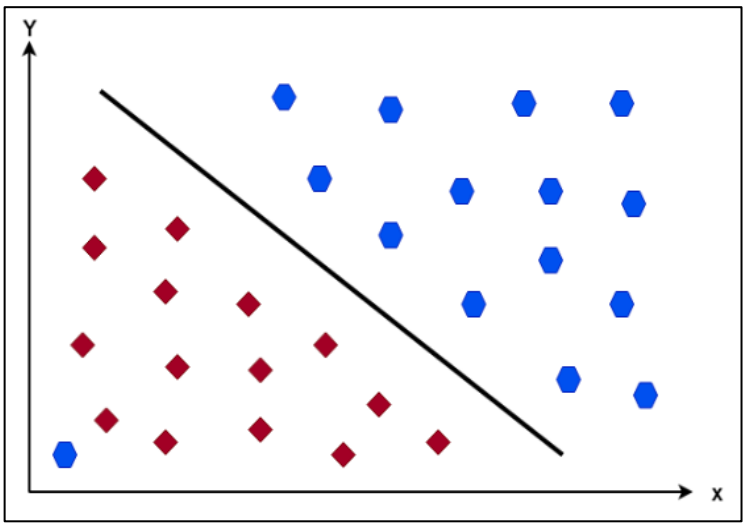



**Figure 4.18: Hyperplane with outlier**

The most problem in real-life is the non-linear type as shown in Figure 4.19. That problem is the structure of vector where main group of vectors inside in another group vector which is very difficult classify to manage. To solve this problem, the Solution has identified.

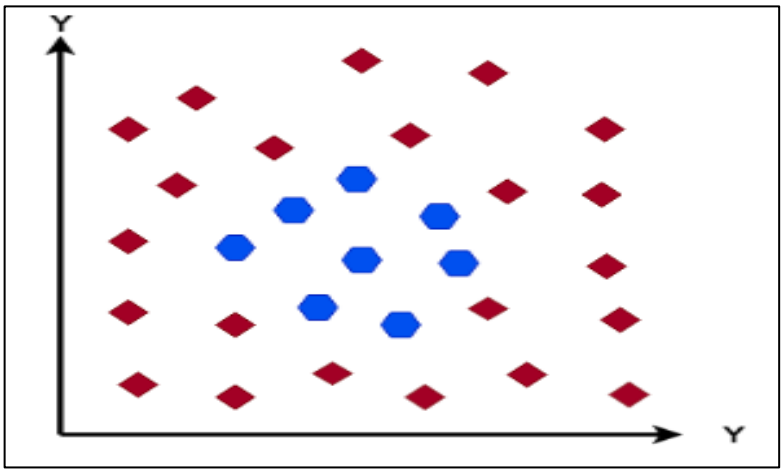



**Figure 4.19: Non-linear graph**

We need to use the kernel function that provides in SVM to solve this problem. The function of the kernel is to transform a non-linear into linear space. It will use the highest dimensional space data point project to enhance the ability to detect the best placement of a hyperplane to separate data by different classes. The kernel function applies data that can be separated into each instance. Typically, two variable data x and y are transformed into several new function spaces defined as z. This is achieved by transforming a two-dimensional spacing into a multi-dimensional spacing shown in figure 4.20.

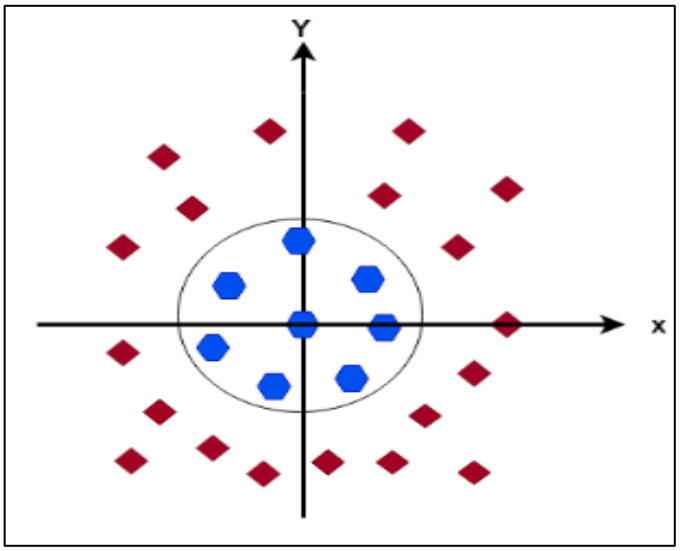



**Figure 4.20: Multi-Dimensional Space**

In addition, it defined as a multi-dimensional space divided into four forms, called Linear, Radial Basis Function (RBF), Polynomial, and Sigmoid. This function is a data point that has its own function for the highest dimensional space by improving the ability to define the best of hyperplane placement to separate data points by splitting it into different classes. Table 4.3 shows the kernel type with its own functionality.

**Table 4.3: Type of Kernel**

| Type of Kernel | Description |
|---|---|
| Linear | By using a straight line, the data point will be separated linearly. Kernels that are good at classifying information in two categories at the same time |
| Polynomial | Commonly works with SVM and a similar vector kernel model. In fact, work with separable nonlinear data. |
| Radial Basis Function (RBF) | To activate a function, the Artificial Neural Network will use a radial base function. The main function is to classify nonlinear data that has many uses including system control, the approximation of function, classification, and prediction of time. |
| Sigmoid | Very popular SVM kernel because of its neural network origin |

To select the best kernel, all of them need to test and compared to the accuracy of their hyperplane. Throughout this project, the high precision is selected and used. The new model focuses more on accuracy identification when the detection of the system occurs. Because of this, this model's result accuracy will be compared to other accuracy generated by using BOW, CS, and IG to make sure that this model is more effective in detecting spam comments on YouTube.

## 4.4 Requirement Analysis

Information on the components involved in system development will be defined in this chapter. Therefore, that criterion will be set out in this part of the project, which is important to ensure that process travel is tentatively planned.

### 4.4.1 Software Requirement

In this project, there is some software that will be used to complete the development of the system will be listed as well as a description of the use of the software. Table 4.4 below shows the system requirements.

**Table 4.4: Software Requirement for the Project**

| Software | Description |
|---|---|
| Windows version 10 | Operating system used in conducting the project. |
| Notepad++++ | Platform used to write Java codes for the system. |
| Weka Application | Platform used to train data. |
| Eclipse IDE Application | Platform used to run and test the codes during the implementation phase. |
| Microsoft Excel 2012 | Software used to do sorting and arrangements of data according to its attributes and instances. |
| Microsoft Word 2013 | Software used in completing the report/documentation of the project. |
| Draw.io online | Software used to construct diagrams. |
| Microsoft Project | Software used in constructing Gantt chart |

### 4.4.2 Hardware Requirement

**Personal Workstation (Laptop)**

Hardware requirements that will be used in the project as a workstation is the laptop. The laptop used in this project is the minimum specifications. Table 4.5 shows the specifications of the laptop.

**Table 4.5: Hardware Requirement of the Project**

| Specification | Description |
|---|---|
| Processor Type | Intel® Pentium ® CPU 2020M |
| Processor Speed | 2.40Ghz |
| Display Resolution | 1366x768 |
| Display Type | HD Display |
| Display Size | 14 |
| Operating System | Windows 10 Pro |
| Operating System Architecture | 64 bits |
| RAM | 10 GB |
| Hard Drive Size | 500 GB |

## 4.5 Conclusion

In conclusion, chapter analysis and design are very important because of misrepresentation as well as relevant information and process parameters involved in the development of the method. The all design involved in this section very important to ensure smooth current projects running. Project design prepared for the show again how this project by showing phase flow system. Before that, preliminary testing will be done before this project starts to get a sample of data set. Finally, the need of software and hardware that will help to carry out the development of this system. The next chapter will explain about the implementation phase to be implemented in accordance with the analysis and design are done in this chapter.

## REFERENCES


Abd, T., & Qaisar, S. (2019). YouTube spam comments detection using Artificial Neural Network. *Journal of Engineering and Applied Sciences*, (January 2018). https://doi.org/10.3923/jeasci.2018.9638.9642

Abdulhamid, S. M., Abd Latiff, M. S., Chiroma, H., Osho, O., Abdul-Salaam, G., Abubakar, A. I., & Herawan, T. (2017). A Review on Mobile SMS Spam Filtering Techniques. *IEEE Access*, *5*(February), 15650–15666. https://doi.org/10.1109/ACCESS.2017.2666785

Ahmad, N., & Habib, M. K. (2010). Analysis of Network Security Threats and Vulnerabilities by Development & Implementation of a Security Network Monitoring Solution. *School of Engineering Department of Telecommunication Blekinge Institute of Technology SE - 371 79 Karlskrona Sweden*.

Ahmed, K., Verma, S., Kumar, N., & Shekbar, J. (2011). Classification Of Internet Security Attacks. *Proceedings of the 5th National Conference on Computing for Nation Development (INDIACom)*, (May), 2–5. https://doi.org/10.7448/IAS.16.1.18445

Aiyar, S., & Shetty, N. P. (2018). N-Gram Assisted Youtube Spam Comment Detection. *Procedia Computer Science*, *132*(Iccids), 174–182. https://doi.org/10.1016/j.procs.2018.05.181

Akanksha, G. (2014). Computer network security issues. *International Research Journal of Management Science & Technology(IRJMST)*, *5*(9), 765–778. https://doi.org/10.1007/978-94-010-0556-2_37

Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2016). TubeSpam: Comment spam filtering on YouTube. *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, (2012), 138–143. https://doi.org/10.1109/ICMLA.2015.37

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. Retrieved from http://arxiv.org/abs/1707.02919

Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G., & Spyropoulos, C. D. (2000). An Evaluation of Naive Bayesian Anti-Spam Filtering. *Machine*

*Learning in the New Information Age*, (September 2014), 9–17. https://doi.org/10.1145/345508.345569

Annadatha, A., & Stamp, M. (2018). Image spam analysis and detection. *Journal of Computer Virology and Hacking Techniques*, *14*(1), 39–52. https://doi.org/10.1007/s11416-016-0287-x

Araujo, L., & Martinez-Romo, J. (2010). Web spam detection: New classification features based on qualified link analysis and language models. *IEEE Transactions on Information Forensics and Security*, *5*(3), 581–590. https://doi.org/10.1109/TIFS.2010.2050767

Attar, A., Rad, R. M., & Atani, R. E. (2013). A survey of image spamming and filtering techniques. *Artificial Intelligence Review*, *40*(1), 71–105. https://doi.org/10.1007/s10462-011-9280-4

Aziz, A., Foozy, C. F. M., Shamala, P., & Suradi, Z. (2018). Youtube spam comment detection using support vector machine and K–nearest neighbor. *Indonesian Journal of Electrical Engineering and Computer Science*, *12*(2), 607–611. https://doi.org/10.11591/ijeecs.v12.i2.pp607-611

Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. (2008). Web Spam Detection: Link-based and Content-based Techniques. *The European Integrated Project Dynamically Evolving Large Scale Information Systems DELIS Proceedings of the Final Workshop*, *222*, 99–113. Retrieved from http://www.chato.cl/papers/becchetti_2008_link_spam_techniques.pdf

Bendovschi, A. (2015). Cyber-Attacks – Trends, Patterns and Security Countermeasures. *Procedia Economics and Finance*, *28*(December 2015), 24–31. https://doi.org/10.1016/s2212-5671(15)01077-1

Bhatia, N., & Vandana. (2010). *Survey of Nearest Neighbor Techniques*. *8*(2), 302–305. Retrieved from http://arxiv.org/abs/1007.0085

Broadhurst, R., & Alazab, M. (2017). Spam and crime. *Regulatory Theory*, 517–532. https://doi.org/10.22459/rt.02.2017.30

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chaudhari, M., & Govilkar, S. (2015). A Survey of Machine Learning Techniques for Sentiment Classification. *International Journal on Computational Science & Applications*, *5*(3), 13–23. https://doi.org/10.5121/ijcsa.2015.5302

Chaudhary, A., Kolhe, S., & Kamal, R. (2013). Machine Learning Classification Techniques : A Comparative Study. *International Journal of Advance Computer Theory Engineeringnal of Advance Computer Theory Engineering*, *2*(4), 2319–2526.

Chaudhry, J. A., Chaudhry, S. A., & Rittenhouse, R. G. (2016). Phishing attacks and defenses. *International Journal of Security and Its Applications*, *10*(1), 247–256. https://doi.org/10.14257/ijsia.2016.10.1.23

Chaudhry, J. A., & Rittenhouse, R. G. (2015). Phishing: Classification and Countermeasures. *2015 7th International Conference on Multimedia, Computer Graphics and Broadcasting (MulGraB)*, (April 2016), 28–31. https://doi.org/10.1109/MulGraB.2015.17

Chen, G., & Chen, J. (2015). A novel wrapper method for feature selection and its applications. *Neurocomputing*, *159*, 219–226. https://doi.org/10.1016/j.neucom.2015.01.070

Chen, Z., & Ji, C. (2009). An information-theoretic view of network-aware malware attacks. *IEEE Transactions on Information Forensics and Security*, *4*(3), 530–541. https://doi.org/10.1109/TIFS.2009.2025847

Chowdury, R., Monsur Adnan, M. N., Mahmud, G. A. N., & Rahman, R. M. (2013). A data mining based spam detection system for YouTube. *8th International Conference on Digital Information Management, ICDIM 2013*, 373–378. https://doi.org/10.1109/ICDIM.2013.6694038

Christina, V., Karpagavalli, S., & Suganya, G. (2010). A Study on Email Spam Filtering Techniques. *International Journal of Computer Applications*, *12*(1), 7–9. https://doi.org/10.5120/1645-2213

Das, M., & Prasad, V. (2014). Analysis of an Image Spam in Email Based on Content Analysis. *International Journal on Natural Language Computing*, *3*(3), 129–140. https://doi.org/10.5121/ijnlc.2014.3313

Dr.S.Kannan, & Gurusamy, V. (2015). *Preprocessing Techniques for Text Mining*.

Dredze, M., Gevaryahu, R., & Elias-Bachrach, A. (2007). Learning fast classifiers for image spam. *4th Conference on Email and Anti-Spam, CEAS 2007*.

Dwivedi, A. K., Mallawaarachchi, I., & Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine*, *36*(14), 2187–2205. https://doi.org/10.1002/sim.7263

Etaiwi, W., & Naymat, G. (2017). The Impact of applying Different Preprocessing Steps on Review Spam Detection. *Procedia Computer Science*, *113*, 273–279. https://doi.org/10.1016/j.procs.2017.08.368

Förstner, W. (2007). Image Preprocessing for Feature Extraction in Digital Intensity, Color and Range Images. *Geomatic Method for the Analysis of Data in the Earth Sciences*, 165–189. https://doi.org/10.1007/3-540-45597-3_4

Genuer, R., Poggi, J. M., Tuleau-Malot, C., & Villa-Vialaneix, N. (2017). Random Forests for Big Data. *Big Data Research*, *9*, 28–46. https://doi.org/10.1016/j.bdr.2017.07.003

Gharatkar, S., Ingle, A., Naik, T., & Save, A. (2018). Review preprocessing using data cleaning and stemming technique. *Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS 2017*, *2018-Janua*, 1–4. https://doi.org/10.1109/ICIIECS.2017.8276011

Google. (2019). YouTube Community Guidelines enforcement At. Retrieved from https://transparencyreport.google.com/youtube-policy/removals

Hamou, R. M., & Amine, A. (2013). The Impact of the Mode of Data Representation for the Result Quality of the Detection and Filtering of Spam. *International Journal of Information Retrieval Research*, *3*(1), 43–59. https://doi.org/10.4018/ijirr.2013010103

Hans, K., Ahuja, L., & K. Muttoo, S. (2014). Approaches for Web Spam Detection. *International Journal of Computer Applications*, *101*(1), 38–44. https://doi.org/10.5120/17655-8467

Hayati, P., & Potdar, V. (2008). Evaluation of spam detection and prevention frameworks for email and image spam. *Proceedings of the 10th International Conference on Information Integration and Web-Based Applications & Services - IiWAS '08*, 520. https://doi.org/10.1145/1497308.1497402

Hayati, P., Potdar, V., Talevski, A., Firoozeh, N., Sarenche, S., & Yeganeh, E. A. (2010). Definition of Spam 2.0: New spamming boom. *4th IEEE International Conference on Digital Ecosystems and Technologies - Conference Proceedings of IEEE-DEST 2010*, *DEST 2010*, (May), 580–584. https://doi.org/10.1109/DEST.2010.5610590

Hema, D. A., & Saravanakumar, R. (2018). A Survey on Feature Extraction Technique in Image Processing. *International Journal of Trend in Scientific Research and Development*, *Volume-2*(Issue-4), 448–451. https://doi.org/10.31142/ijtsrd12937

Issac, B. (2010). Spam detection approaches with case study implementation on spam corpora. *Cases on ICT Utilization, Practice and Solutions: Tools for Managing Day-to-Day Issues*, (November 2010), 194–212. https://doi.org/10.4018/978-1-60960-015-0.ch012

Jiansheng, W., & Tao, D. (2008). Research in anti-spam method based on bayesian filtering. *Proceedings - 2008 Pacific-Asia Workshop on Computational Intelligence and Industrial Application, PACIIA 2008*, *2*, 887–891. https://doi.org/10.1109/PACIIA.2008.180

Kadhim, A. I., Cheah, Y. N., & Ahamed, N. H. (2015). Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering. *Proceedings - 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, ICAIET 2014*, 69–73. https://doi.org/10.1109/ICAIET.2014.21

Kanaris, I., Kanaris, K., Houvardas, I., & Stamatatos, E. (2006). *Words vs. character n-grams for anti-spam filtering*. *XX*(X), 1–20.

Kang, M., Ahn, J., & Lee, K. (2017). Opinion mining using ensemble text hidden Markov models for text classification. *Expert Systems with Applications*, *94*, 218–227. https://doi.org/10.1016/j.eswa.2017.07.019

Karami, A., & Zhou, L. (2014). Improving static SMS spam detection by using new content-based features. *20th Americas Conference on Information Systems, AMCIS 2014*, 1–9.

Ketari, L. M., Chandra, M., & Khanum, M. A. (2012). A study of image spam filtering techniques. *Proceedings - 4th International Conference on Computational Intelligence and Communication Networks, CICN 2012*, 245–250. https://doi.org/10.1109/CICN.2012.34

Khan, A. A. (2013). Preventing Phishing Attacks using One Time Password and User Machine Identification. *International Journal of Computer Applications (0975, 68*(3), 7–11.

Khawandi, S., Abdallah, F., & Ismail, A. (2019). *A Survey On Image Spam Detection Techniques*. (January), 13–27. https://doi.org/10.5121/csit.2019.90102

Kim, S., Chang, H., Lee, S., Yu, M., & Kang, J. (2015). Deep semantic frame-based deceptive opinion spam analysis. *International Conference on Information and Knowledge Management, Proceedings*, *19-23-Oct-*, 1131–1140. https://doi.org/10.1145/2806416.2806551

Konakalla, A., & Veeranki, B. (2013). Evolution of Security Attacks and Security Technology. *Ijcsmc*, *2*(11), 270–276.

Kondakci, S. (2008). Epidemic state analysis of computers under malware attacks. *Simulation Modelling Practice and Theory*, *16*(5), 571–584. https://doi.org/10.1016/j.simpat.2008.02.011

Krishnamurthy, V. (2015). *Internet spam threats and email exploitation – A scuffle with inbox attack*. (January 2014). https://doi.org/10.6088/ijaser.030400015

Kumar, M., Babaeizadeh, M., Erhan, D., Finn, C., Levine, S., & Dinh, L. (2015). *VideoFlow : A Flow-Based Generative Model for Video*.

Kumar Sharma, A., Kumar Prajapat Assistant Professor, S., & Aslam, M. (2014). A Comparative Study between Naïve Bayes and Neural Network (MLP) Classifier for Spam Email Detection. *International Journal of Computer Applications® (IJCA)*, 975–8887.

Larroza, A., Moratal, D., Paredes-Sánchez, A., Soria-Olivas, E., Chust, M. L., Arribas, L. A., & Arana, E. (2015). Support vector machine classification of brain metastasis and radiation necrosis based on texture analysis in MRI. *Journal of Magnetic Resonance Imaging*, *42*(5), 1362–1368. https://doi.org/10.1002/jmri.24913

Li, Y., Nie, X., & Huang, R. (2018). Web spam classification method based on deep belief networks. *Expert Systems with Applications*, *96*, 261–270. https://doi.org/10.1016/j.eswa.2017.12.016

Liu, Q., Zhang, F. L., Qin, Z. G., Wang, C., Chen, S., & Ma, Q. M. (2010). Feature selection for image spam classification. *2010 International Conference on Communications, Circuits and Systems, ICCCAS 2010 - Proceedings*, 294–297. https://doi.org/10.1109/ICCCAS.2010.5581994

M.Nazreen, & S.Munawara. (2013). A Comprehensive Study of Phishing Attacks. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *4*(6), 783–786.

Ma, Q. (2013). The process and characteristics of phishing attacks - A small international trading company case study. *Journal of Technology Research*.

Mahmoud, T. M., Mahfouz, A. M., Minia, E., & Minia, E. (2012). SMS Spam Filtering Technique Based on Artificial Immune System. *International Journal of Computer Science Issues*, *9*(2), 589–597.

Mathew, N. V., & Ramani Bai, V. (2017). Analyzing the Effectiveness of N-gram

Technique Based Feature Set in a Naive Bayesian Spam Filter. *Proceedings of IEEE International Conference on Emerging Technological Trends in Computing, Communications and Electrical Engineering, ICETT 2016*. https://doi.org/10.1109/ICETT.2016.7873648

Mehta, B., Nangia, S., Gupta, M., & Nejdl, W. (2008). Detecting image spam using visual features and near duplicate detection. *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, 497–506. https://doi.org/10.1145/1367497.1367565

Mishra, B. K., & Ansari, G. M. (2012). Differential epidemic model of virus and worms in computer network. *International Journal of Network Security*, *14*(3), 149–155. https://doi.org/10.6633/IJNS.201205.14(3).03

Mohammad, A. H. (2019). Arabic Text Classification: A Review. *Modern Applied Science*, *13*(5), 88. https://doi.org/10.5539/mas.v13n5p88

MyCERT. (2017). *Reported Incidents based on General Incident Classification Statistics 2018*. Retrieved from https://www.mycert.org.my/

MyCERT. (2018). *Reported Incidents based on General Incident Classification Statistics 2018*. Retrieved from https://www.mycert.org.my/

MyCERT. (2019). *Reported Incidents based on General Incident Classification Statistics 2019*. Retrieved from https://www.mycert.org.my/

Mylonas, A., Dritsas, S., Tsoumas, B., & Gritzalis, D. (2012). On the Feasibility of malware attacks in smartphone platforms. *Communications in Computer and Information Science*, *314*, 217–232. https://doi.org/10.1007/978-3-642-35755-8_16

Nath, A., & Dasgupta, A. (2016). Classification of Machine Learning Algorithms. *International Journal of Innovatice Research in Advanced Engineering*, *3*(March), 6–11.

Nayak, J., Naik, B., & Behera, H. S. (2015). A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges. *International Journal of Database Theory and Application*, *8*(1), 169–186. https://doi.org/10.14257/ijdta.2015.8.1.18

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015). Hidden Markov models for cancer classification using gene expression profiles. *Information Sciences*, *316*(September), 293–307. https://doi.org/10.1016/j.ins.2015.04.012

Nisioi, S., Bucur, A., & P.Dinu, L. (2018). Lexical Analysis and Content Extraction

from Customer-Agent Interactions. *Human Language Technologies Research Center*, 132–136.

Paetzold, G. H., & Specia, L. (2017). A Survey on Lexical Simplification. *Journal of Artificial Intelligence Research*, *60*, 549–593. https://doi.org/10.1613/jair.5526

Pawar, M. V., & Anuradha, J. (2015). Network security and types of attacks in network. *Procedia Computer Science*, *48*(C), 503–506. https://doi.org/10.1016/j.procs.2015.04.126

Perumal, S., & Velmurugan, T. (2018). *Preprocessing by Contrast Enhancement Techniques for Medical Images*. *118*(18), 3681–3688.

Puspita, R. H., & Rohedi, D. (2018). The Impact of Internet Use for Students. *IOP Conference Series: Materials Science and Engineering*, *306*(1). https://doi.org/10.1088/1757-899X/306/1/012106

Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *Eurasip Journal on Advances in Signal Processing*, *2016*(1). https://doi.org/10.1186/s13634-016-0355-x

Ramesh, N. (2013). *a Survey of Different Types of Network Security*. 28–31.

Rana, M. I., Khalid, S., & Akbar, M. U. (2015). News classification based on their headlines: A review. *17th IEEE International Multi Topic Conference: Collaborative and Sustainable Development of Technologies, IEEE INMIC 2014 - Proceedings*, 211–216. https://doi.org/10.1109/INMIC.2014.7097339

Rao, A. S., Avadhani, P. ., & Chaudhuri, N. B. (2016). A Content-Based Spam E-Mail Filtering Approach Using Multilayer Percepton Neural Networks. *International Journal of Engineering Trends and Technology*, *41*(1), 44–45. https://doi.org/10.14445/22315381/ijett-v41p210

Rao, J. M., & Reiley, D. H. (2012). The economics of spam. *Journal of Economic Perspectives*, *26*(3), 87–110. https://doi.org/10.1257/jep.26.3.87

Reddy, K. S., & Reddy, E. S. (2019). Integrated approach to detect spam in social media networks using hybrid features. *International Journal of Electrical and Computer Engineering (IJECE)*, *9*(1), 562. https://doi.org/10.11591/ijece.v9i1.pp562-569

Roy, S. S., & Viswanatham, V. M. (2016). Classifying spam emails using artificial intelligent techniques. *International Journal of Engineering Research in Africa*, *22*(February), 152–161. https://doi.org/10.4028/www.scientific.net/JERA.22.152

Rudd, E. M., Rozsa, A., Günther, M., & Boult, T. E. (2017). A Survey of Stealth Malware Attacks, Mitigation Measures, and Steps Toward Autonomous Open World Solutions. *IEEE Communications Surveys and Tutorials*, *19*(2), 1145–1172. https://doi.org/10.1109/COMST.2016.2636078

S.Mangrulkar, N., R. Bhagat Patil, A., & S. Pande, A. (2014). Network Attacks and Their Detection Mechanisms: A Review. *International Journal of Computer Applications*, *90*(9), 37–39. https://doi.org/10.5120/15606-3154

Samsudin, N. M., Mohd Foozy, C. F. B., Alias, N., Shamala, P., Othman, N. F., & Wan Din, W. I. S. (2019). Youtube spam detection framework using naïve bayes and logistic regression. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1508–1517. https://doi.org/10.11591/ijeecs.v14.i3.pp1508-1517

Saritas, M. M. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88–91. https://doi.org/10.18201/ijisae.2019252786

SECURELIST. (2019). *Spam and phishing in Q2 2019*. *2019*. Retrieved from website: https://securelist.com/spam-and-phishing-in-q2- 2019/92379/

Seeja, K. R., & Zareapoor, M. (2015). Feature Extraction or Feature Selection for Text Classification : A Case Study on Phishing Email Detection. *I.J. Information Engineering and Electronic Business*, *2*(March), 60–65. https://doi.org/10.5815/ijieeb.2015.02.08

Shameena, N., & Jabbar, R. (2014). Techniques on Cardiac Medical Images. *International Journal of Engineering Research and Technology (IJERT)*, *3*(4), 336–341. https://doi.org/10.1177/1747016117711971

Sharmin, S., & Zaman, Z. (2018). Spam detection in social media employing machine learning tool for text mining. *Proceedings - 13th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2017*, *2018-Janua*, 137–142. https://doi.org/10.1109/SITIS.2017.32

Sivakumar, A. (2017). A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining. *International Journal of Pure and Applied Mathematics*, *117*(20), 785–794.

Spirin, N., & Han, J. (2012). Survey on Web Spam Detection: Principles and Algorithms. *SIGKDD Explor. Newsl.*, *13*(2), 50–64.

https://doi.org/10.1145/2207243.2207252

Subramanya, S. R., & Lakshminarasmhan, N. (2001). Computer Viruses. *IEEE Potentials*, Vol. 20, pp. 16–19. https://doi.org/10.1109/45.969588

Suganya, V. (2016). A Review on Phishing Attacks and Various Anti Phishing Techniques. *International Journal of Computer Applications*, *139*(1), 20–23.

Taylor, P., Jin, C., Ma, T., Hou, R., Tang, M., Tian, Y., … Al-rodhaan, M. (2015). *Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization*. (May), 37–41. https://doi.org/10.1080/03772063.2015.1021385

Tewari, A., & Jangale, S. (2016). Spam Filtering Methods and machine Learning Algorithm - A Survey. *International Journal of Computer Applications*, *154*(6), 8–12. https://doi.org/10.5120/ijca2016912153

Thaseen, I. S., & Kumar, C. A. (2016). *Intrusion Detection Model Using Chi Square Feature Selection and Modified Naïve Bayes Classifier*. (November 2018). https://doi.org/10.1007/978-3-319-30348-2

Tilve, A. K. S., & Jain, S. N. (2017). A SURVEY ON MACHINE LEARNING TECHNIQUES FOR TEXT CLASSIFICATION. *A SURVEY ON MACHINE LEARNING TECHNIQUES FOR TEXT CLASSIFICATION*, *6*(2), 513–520. Retrieved from http: // www.ijesrt.com Fig.1.

Tiwari, A. (2014). Introduction to Network Security , Attacks and Services. *Computer Science And Enginerring , RKDF University Bhopal M.P.*, *3*(4), 1–15. https://doi.org/10.1016/j.cell.2009.01.043

Tran, H. (2019). *A SURVEY OF MACHINE LEARNING AND DATA MINING TECHNIQUES USED IN MULTIMEDIA SYSTEM*. (113), 13–21. https://doi.org/10.13140/RG.2.2.20395.49446/1

Trivedi, S. K. (2016). A study of machine learning classifiers for spam detection. *2016 4th International Symposium on Computational and Business Intelligence, ISCBI 2016*, (September 2016), 176–180. https://doi.org/10.1109/ISCBI.2016.7743279

Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2013). The impact of feature extraction and selection on SMS spam filtering. *Elektronika Ir Elektrotechnika*, *19*(5), 67–72. https://doi.org/10.5755/j01.eee.19.5.1829

Uysal, Alper Kursat. (2018). Feature Selection for Comment Spam Filtering on YouTube. *DATA SCIENCE AND APPLICATIONS*, *1*(1).

Uysal, Alper Kursat, & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, *50*(1), 104–112. https://doi.org/10.1016/j.ipm.2013.08.006

Vani, K., & Gupta, D. (2014). Using K-means cluster based techniques in external plagiarism detection. *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, 1268–1273. https://doi.org/10.1109/IC3I.2014.7019659

Verbert, K., Manouselis, N., Drachsler, H., & Duval, E. (2012). Dataset-driven research to support learning and knowledge analytics. *Educational Technology and Society*, *15*(3), 133–148.

Wan, C. H., Lee, L. H., Rajkumar, R., & Isa, D. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. *Expert Systems with Applications*, *39*(15), 11880–11888. https://doi.org/10.1016/j.eswa.2012.02.068

Wang, B., Zubiaga, A., Liakata, M., & Procter, R. (2015). Making the most of tweet-inherent features for social spam detection on twitter. *CEUR Workshop Proceedings*, *1395*, 10–16.

Wattenhofer, M., Wattenhofer, R., & Zhu, Z. (2012). The you tube social network. *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 354–361.

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2008). *ESFS : A new embedded feature selection method based on SFS*. (May 2014).

Yogatama, D., Dyer, C., Ling, W., & Blunsom, P. (2017). *Generative and Discriminative Text Classification with Recurrent Neural Networks*.

Yusof, Y., & Sadoon, O. H. (2017). Detecting Video Spammers in Youtube Social Media. *ICOCI Kuala Lumpur. Universiti Utara Malaysia*, (082), 25–27. Retrieved from http://www.uum.edu.my

Zhang, N., & Lu, W. F. (2007). An efficient data preprocessing method for mining customer survey data. *IEEE International Conference on Industrial Informatics (INDIN)*, *1*, 573–578. https://doi.org/10.1109/INDIN.2007.4384821

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006

Zhang, Z., & Gupta, B. B. (2018). Social media security and trustworthiness:

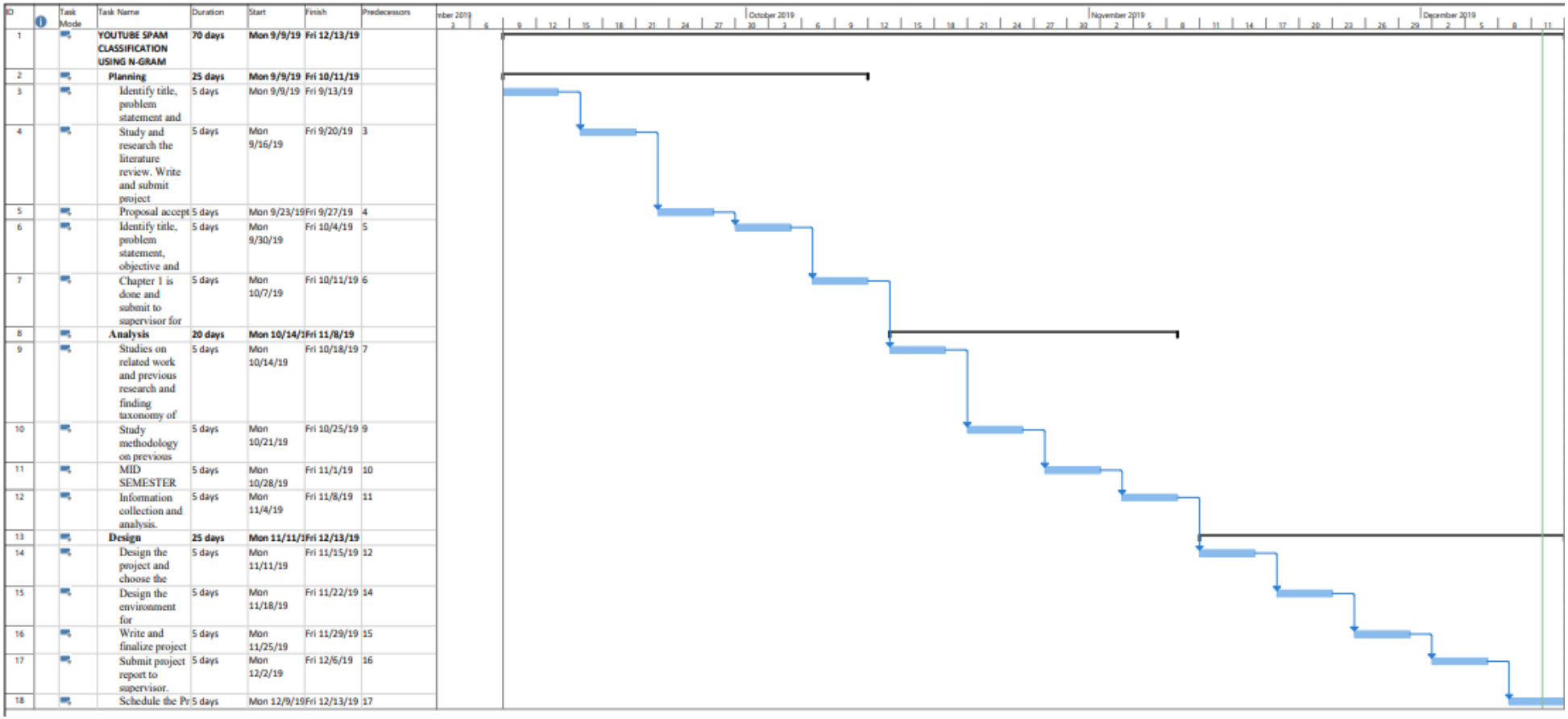
Overview and new direction. *Future Generation Computer Systems*, *86*, 914–925. https://doi.org/10.1016/j.future.2016.10.007

Zhu, Z., Ong, Y., & Dash, M. (2007). *Wrapper-Filter Feature Selection Algorithm Using A Memetic Framework*. (May 2014). https://doi.org/10.1109/TSMCB.2006.883267

Zorarpacı, E., & Ay, S. (2016). *A hybrid approach of differential evolution and artificial bee colony for feature selection*. *62*, 91–93.

# APPENDIX

A. Gantt Chart part



| ID | Task Mode | Task Name | Duration | Start | Finish | Predecessors |
|---|---|---|---|---|---|---|
| 1 | | **YOUTUBE SPAM CLASSIFICATION USING N-GRAM** | **70 days** | **Mon 9/9/19** | **Fri 12/13/19** | |
| 2 | | **Planning** | **25 days** | **Mon 9/9/19** | **Fri 10/11/19** | |
| 3 | | Identify title, problem statement and | 5 days | Mon 9/9/19 | Fri 9/13/19 | |
| 4 | | Study and research the literature review. Write and submit project | 5 days | Mon 9/16/19 | Fri 9/20/19 | 3 |
| 5 | | Proposal accept | 5 days | Mon 9/23/19 | Fri 9/27/19 | 4 |
| 6 | | Identify title, problem statement, objective and | 5 days | Mon 9/30/19 | Fri 10/4/19 | 5 |
| 7 | | Chapter 1 is done and submit to supervisor for | 5 days | Mon 10/7/19 | Fri 10/11/19 | 6 |
| 8 | | **Analysis** | **20 days** | **Mon 10/14/1** | **Fri 11/8/19** | |
| 9 | | Studies on related work and previous research and finding taxonomy of | 5 days | Mon 10/14/19 | Fri 10/18/19 | 7 |
| 10 | | Study methodology on previous | 5 days | Mon 10/21/19 | Fri 10/25/19 | 9 |
| 11 | | MID SEMESTER | 5 days | Mon 10/28/19 | Fri 11/1/19 | 10 |
| 12 | | Information collection and analysis. | 5 days | Mon 11/4/19 | Fri 11/8/19 | 11 |
| 13 | | **Design** | **25 days** | **Mon 11/11/1** | **Fri 12/13/19** | |
| 14 | | Design the project and choose the | 5 days | Mon 11/11/19 | Fri 11/15/19 | 12 |
| 15 | | Design the environment for | 5 days | Mon 11/18/19 | Fri 11/22/19 | 14 |
| 16 | | Write and finalize project | 5 days | Mon 11/25/19 | Fri 11/29/19 | 15 |
| 17 | | Submit project report to supervisor. | 5 days | Mon 12/2/19 | Fri 12/6/19 | 16 |
| 18 | | Schedule the Pr | 5 days | Mon 12/9/19 | Fri 12/13/19 | 17 |