IMAGE SPAM DETECTION USING FREQUENT ITEM MINING TECHNIQUE

NOR ANIS SYAZANA BINTI ZULKIFLI

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

JUDUL: IMAGE SPAM DETECTION USING FREQUENT ITEM MINING TECHNIQUE

SESI PENGAJIAN: 2016/2017

Saya NOR ANIS SYAZANA BINTI ZULKIFLI mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/)

_____ SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

_____ TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

_____ TIDAK TERHAD

(TANDATANGAN PENULIS)

Alamat tetap: NO 155 JALAN
MERANTI 1, TAMAN JERAI
FASA 2, 08300 GURUN, KEDAH

Tarikh: 25/8/2017

(TANDATANGAN PENYELIA)

NOR AZMAN BIN MAT ARIFF

Nama Penyelia

Tarikh: 25/8/2017

CATATAN: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)
** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

IMAGE SPAM DETECTION USING FREQUENT ITEM MINING TECHNIQUE

NOR ANIS SYAZANA BINTI ZULKIFLI

This report is submitted in partial fulfillment of the requirements for the Bachelor of Computer Science (Computer Networking)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2017

DECLARATION

I hereby declare that this project report entitled

IMAGE SPAM DETECTION USING FREQUENT ITEM MINING TECHNIQUE

is written by me and is my own effort and that no part has been plagiarized

without citations.

STUDENT : Anis Date : 25/8/2017

(NOR ANIS SYAZANA BINTI ZULKIFLI)

SUPERVISOR : Date : 25/8/2017

(EN. NORAZMAN BIN MAT ARIFF)

DEDICATION

This final project is dedicated to my beloved parents Mr. Zulkifli BIn Yahya and Mrs. Salmiah Binti Md Ali for their endless support and helps throughout the year in completing my studies and always pray the best for me.

I also dedicate this dissertation to my friends who have supported me throughout this project. I will always appreciate all they have done, especially to Sarah Syamimi and Amizah Aida for helping me from FYP 1 until FYP 2.

Special dedication to my supervisor, Encik Azman Bin Mat Ariff who has guided and never get tired in supports, helping and motivating me while making the progress for this project.

Last and but not least, do not forget to my evaluator, P.M. Dr. Faizal Bin Abdollah for his positive advices and feedback about this project,

.

ACKNOWLEDGEMENT

First and foremost, I would like to express my thanks to Allah S.W.T for giving me strength for completing the whole process of this project. Without His blessings, I am sure this project cannot complete.

Special thanks to my supervisor, Enick Azman Bin Mat Ariff for his guidance, constant supervision and kindness in completing this project. Without his guide, I am surely would not know the right trail.

I would also like to thanks to my family members who always support and pray the best for me and constantly give me motivation throughout my project. Besides, thanks to my colleague who have been very helpful during project development.

Thank you.

ABSTRACT

Spam is usually sent in the form of a text message in which specific words in the message can be used by spam blocking software to prevent the message from reaching our Inbox. With image spamming, the text is placed inside the image in an effort to bypass the spam blocking software. Since images are considered a normal part of a recipient's email message and the spam blocking software is mainly designed for text, the spammer is successful in getting the message to reach our Inbox. One of the previous study has used frequent item mining technique in order to extract features. However, the researcher only considered minimum weightage scheme for feature weight assignment. Thus, the main objective of this project is to generate feature vectors using weightage schemes, which is a maximum weightage scheme assignments. For further investigation, an ensemble method also is applied for weightage schemes. Firstly, sift descriptor is used to represent an image. Sift keypoint make the process of clustering and each of keypoint were collected the cluster number. Bag-of-word feature vectors are generated directly. Then, the frequent items are generated from BOW feature vector using of weightage schemes. SVM classifier is used as image spam classifier to train and produce a models for scheme. Lastly, an ensemble method used for models to obtain the best models. The significant contribution for this project is using the weightage schemes of maximum of the frequent item mining (FIM) technique to generate a feature vectors that is capable of detecting image with better.

ABSTRAK

Spam biasanya dihantar dalam bentuk mesej teks di mana kata-kata khusus dalam mesej itu boleh digunakan oleh perisian menyekat spam untuk menghalang mesej daripada mencapai Peti Masuk kami. Dengan spamming imej, teks diletakkan di dalam imej dalam usaha untuk memintas perisian menghalang spam. Oleh kerana imej dianggap sebagai sebahagian daripada mesej e-mel penerima dan perisian penyekatan spam terutamanya direka untuk teks, spammer berjaya mendapatkan mesej untuk mencapai Peti Masuk kami. Salah satu kajian terdahulu telah menggunakan teknik perlombongan item kerap untuk mengekstrak ciri-ciri. Walau bagaimanapun, penyelidik hanya menganggap skema weightage minimum untuk tugasan berat ciri. Oleh itu, matlamat utama projek ini adalah untuk menghasilkan vektor ciri menggunakan skema weightage, yang merupakan tugasan skema pemberat weight maksimum. Untuk siasatan lanjut, kaedah ensemble juga digunakan untuk skema weightage. Pertama, penyaring deskriptor digunakan untuk mewakili imej. Menapis keypoint membuat proses clustering dan setiap keypoint dikumpulkan nombor kluster. Vektor ciri beg-perkataan dihasilkan secara langsung. Kemudian, item kerap dijana daripada vektor ciri BOW menggunakan skema weightage. Pengelas SVM digunakan sebagai pengelas imej spam untuk melatih dan menghasilkan model untuk skema. Akhir sekali, kaedah ensemble yang digunakan untuk model untuk mendapatkan model terbaik. Sumbangan besar untuk projek ini menggunakan skema berat maksimum teknik perlombongan item kerap (FIM) untuk menghasilkan vektor ciri yang mampu mengesan imej dengan lebih baik.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLE

CHAPTER 1

INTRODUCTION

1.1 Introduction

In spite of the fact that a considerable measure of programming has been created as of late to square spam from your email customer, there is another type of spam that is intended to get around spam blockers and into your Inbox. The new type of spamming is known as picture spamming and uses pictures rather than content to sidestep spam blocking programming.

Spam is normally sent as an instant message in which particular words in the message can be utilized by spam blocking programming to keep the message from achieving your Inbox. With picture spamming, the content is put inside the picture with an end goal to sidestep the spam blocking programming.

Since pictures are viewed as a typical piece of a beneficiary's email message and the spam blocking programming is for the most part intended for content, the spammer is fruitful in getting the message to come to your Inbox.

Spam channels are intended to peruse particular words that decide if the spam ought to be sifted through. The messages are then sent to a spam envelope where the beneficiary can see them or not.

Later spam channels have fixed security by taking out the spam messages out and out and enhancing the criteria that decides a message as spam. Accordingly, spammers thought of an approach to consolidate content with pictures to get around the fixed security of spam channels.

1.2 Problem Statement

The problem statement that has been identified is summarized in table 1.1 below

Table 1.1 Summaries of Problem Statement

PS	Problem Statement
PS1	The effect of image spam

PS1: The effect of image spam.

The examples of the problem is bandwidth used for free and the email users will waste time to delete spam emails or the image spam will attach for virus and also malware.

1.3 Project Question

In this research, there is three Project Question (PQ) that needs to be answered in this project. The summary of project question is shown in Table 1.2

Table 1.2 Summaries of Project Question

PQ	Project Questions
PQ1	What are the purposes on this project?
PQ2	Which method is the best to detect an image spam?
PQ3	What is the method that used in this project?

PQ1: What are the purposes on this project?

To investigate the performance of the classifier using the feature vector generated weightage schemes and maximum.

PQ2: Which method is the best to detect an image spam?

To indentify the performance which is the best between a single classifier.

PQ3: What is the method that used in this project?

To identify the performance of ensemble method when combined of these models.

1.4 Project Objective

Based on this research, there are three Project Objectives (PO) that are developed as follows in table 1.3.

Table 1.1 Summaries of Project Objectives

PQ	Project Objectives
PO1	To investigate the performance of the classifier using the feature vector generated from weightage schemes, and maximum.
PO2	To indentify the performance which is the best between a single classifier
PO3	To indentify the performance of ensemble method when combined of these models.

1.5 Scopes

The scope of this projects is to detect an image spam by using Frequent Itemset Mining Techniques. The main goal of the project is to protect the image of the attack by spam. When image spam is filtered, it can be more secured.

1.6 Project Contribution

Image spam sifting used to guarantees messages you do need endure. Spam keeps vital and true blue correspondence from contacting the target group. Spam separating hinders the garbage and keeps it from hitting the inbox in any case. At that point it can aggregates with information security and email controls since Spam sifting helps organizations with staying consistent and a la mode on security.

By doing this project, I can know what a suitable technique that can detect an image spam. Besides, an image spam also can classify by the dataset. In this case, it can classify an image spam to build a graph or table to show the result after do the experiment.

1.7 Thesis Organization

Chapter 1: Introduction

In this chapter it will discuss the background of project. It also include the problem statement, project question, project scope, project contribution, thesis organization and conclusion.

Chapter 2: Literature Review

In this chapter it will discuss about previous work that are related to the project and review of previous researcher. This chapter also include the solution for the project.

Chapter 3: Project Methodology

In this chapter it will discuss about the methodology that will be used in these projects. This chapter also will explain on the methodology for every step that has been used and the project milestone.

Chapter 4: Analysis and Design

In this chapter it will discuss about the analysis of the project and the design of the project that had been used in the experiment.

1.8 Conclusion

The conclusion for this project is using the two (2) weightage which is Bag Of Word and Minimum schemes and the Maximum of the frequent item mining (FIM) technique to generate a feature vectors that is capable of detecting image with better.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Image spam is a sort of spam, or rather, a spamming technique, in which a spam message is conveyed as an image. This is done trying to circumvent spam filters that sweep for specific catchphrases

Image spam is junk email that replaces content with pictures as a methods for tricking spam channels. Picture conveyance works by implanting code in a HTML message that connections to a picture record on the Web. Image spam is a bigger deplete on system assets than content spam since picture documents are bigger than ASCII character strings. Bigger records require more data transmission and, as a result, cause more prominent corruption of exchange rates.

2.2 Related Work

In this section it will explains about the work that related with the project. It includes all the information about spam, image spam, email spam, web spam, sms spam, social spam, the statistic of cyber crime, machine learning and the related work and others that relate with the project.

2.2.1 What is Spam?

Spam is thought to be electronic garbage mail or garbage newsgroup postings. A few people characterize spam much more for the most part as any unsolicited email. Be that as it may, if a departed sibling finds your email address and sends you a message, this could barely be called spam, despite the fact that it is spontaneous. Genuine spam is for the most part email publicizing for some item sent to a mailing list or newsgroup.

"Spam" is an acronym gotten from the words "spiced" and 'ham'. It's superfluous or spontaneous messages sent over the Internet, regularly to a substantial number of clients, for the motivations behind promoting, phishing, spreading malware, and so forth.

In view of past specialist, There exist different meanings of what spam (likewise called garbage mail) is and how it varies from true blue mail (additionally called non-spam, real mail or ham). The most limited among the well known definitions describes spam as "spontaneous mass email" (Androutsopoulos et al. 2000b; SPAMHAUS 2005). In some cases the word business is included, however this expansion is far from being obviously true. The TREC Spam Track depends on a comparative definition: spam is "spontaneous, undesirable email that was sent aimlessly, specifically or in a roundabout way, by a sender having no present association with the client" (Cormack and Lynam 2005a). Another generally acknowledged definition expresses that "Web spam is at least one spontaneous messages, sent or posted as a major aspect of a bigger gathering of messages, all having generously indistinguishable substance" (SpamDefined 2001). Coordinate Marketing Association proposed to utilize "spam" just for messages with specific sorts of substance, for example, obscenity, however this thought met no excitement, being viewed as an endeavor to legitimize different sorts of spam (SPAMHAUS 2003).

As should be obvious, the basic point is that spam is spontaneous, as per a generally refered to equation "spam is about assent, not content" (SPAMHAUS 2005). It is important to say that the thought of being spontaneous is difficult to catch. Actually, regardless of the wide concession to this kind of definitions the channels need to depend on substance and methods for conveyance of messages to perceive spam from honest to goodness mail. Among the most recent work it is fascinating to specify Zinman and Donath (2007), who still like to depend on substance and a client's close to home judgment to characterize spam.

2.2.1.1 What is Image Spam?

Spam is generally sent as an instant message in which particular words in the message can be utilized by spam blocking programming to keep the message from achieving your Inbox. With picture spamming, the content is set inside the picture with an end goal to sidestep the spam blocking programming.

Since pictures are viewed as a typical piece of a beneficiary's email message and the spam blocking programming is essentially intended for content, the spammer is effective in getting the message to come to your Inbox. Figure 2.1 shows the example of image spam

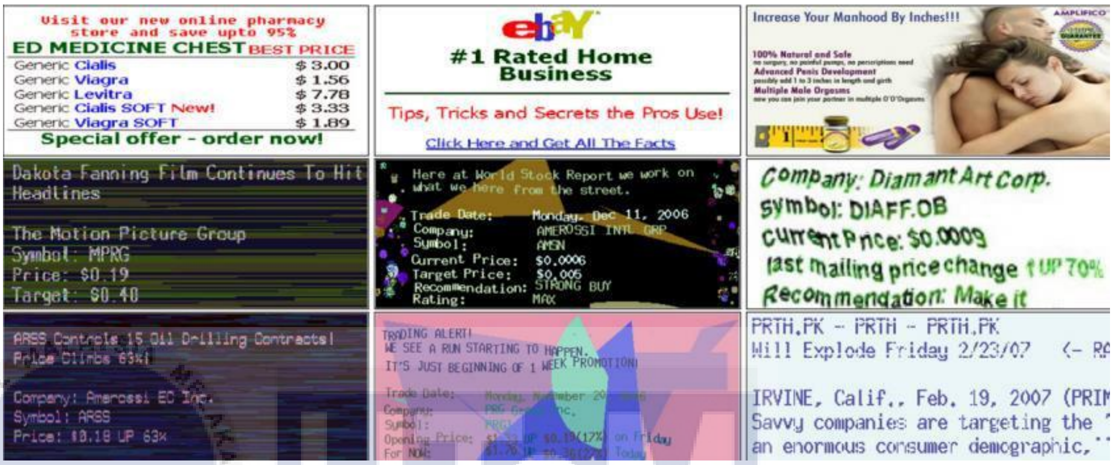

Visit our new online pharmacy store and save upto 95%
ED MEDICINE CHEST BEST PRICE
Generic Cialis $ 3.00
Generic Viagra $ 1.56
Generic Levitra $ 7.78
Generic Cialis SOFT New! $ 3.33
Generic Viagra SOFT $ 1.89
Special offer - order now!
ebay
#1 Rated Home Business
Tips, Tricks and Secrets the Pros Use!
Click Here and Get All The Facts
Increase Your Manhood By Inches!!!
AMPLIFICO
100% Natural and Safe
Advanced Penis Development
Multiple Male Orgasms
Dakota Fanning Film Continues To Hit Headlines
The Motion Picture Group
Symbol: MPRG
Price: $0.19
Target: $0.40
Here at World Stock Report we work on what we here from the street.
Trade Date: Monday, Dec 11, 2006
Company: AMEROSSI INTL GRP
Symbol: AMSN
Current Price: $0.0006
Target Price: $0.005
Recommendation: STRONG BUY
Rating: MAX
Company: Diamant Art Corp.
symbol: DIAFF.OB
current Price: $0.0009
last mailing price change ↑UP 70%
Recommendation: Make it
ARSS Controls 15 Oil Drilling Contracts!
Price Climbs 63%!
Company: Amerossi EC Inc.
Symbol: ARSS
Price: $0.18 UP 63%
TRADING ALERT!
WE SEE A RUN STARTING TO HAPPEN.
IT'S JUST BEGINNING OF 1 WEEK PROMOTION!
Trade Date:
Company:
Symbol:
Opening Price:
For NOW:
PRTH.PK - PRTH - PRTH.PK
Will Explode Friday 2/23/07 <- RA
IRVINE, Calif., Feb. 19, 2007 (PRIM
Savvy companies are targeting the
an enormous consumer demographic,

Figure 2.1 : Example of real spam images

In past analyst, picture spam implies an email which contains picture does not really contain picture spam, while picture spam contains picture. Picture spam is another origination, so meaning of picture spam is not uniform in principle up to now.

As far as anyone is concerned, picture spam implies that the genuine spam message is moved into a picture appended to the message in some logical written works [1-3]. In any case, the definition is not finished, which portrays the piece of the attributes of picture spam. (Jurnal : A Simple Method for Filtering Image Spam)

In light of past specialist, The picture spam, in which the message content of the spam is displayed as pictures in a picture document combined with including commotions in the photos and utilizing muddling system, is imagined to go around those separating programming devoted to content based space sifting. Ordinarily, the measure of a picture spam email is around 3 to 4 times bigger than a relating plain content based email. This element carries with a few direct damages two of which that can be instantly seen by instinct are the dispute of data transmission when exchanging picture spam over the Internet and the additional prerequisite for storage room.

2.2.1.2 What is Email Spam?

Email spam is any email that meets the accompanying three criteria which is Anonymity, Mass Mailing and Unsolicited. Secrecy is the address and character of the sender are hidden. Next, mass mailing is the email is sent to substantial gatherings of individuals. Last, spontaneous is the email is not asked for by the beneficiaries

So, spam email is any email that was not asked for by a client but rather was sent to that client and numerous others, regularly (yet not generally) with malevolent expectation. The source and personality of the sender is mysterious and there is no alternative to stop getting future messages.

2.2.1.2.1 How to prevent Email Spam?

While getting some spam might be unavoidable, clients can decrease the sum that makes it into their inbox. Most email customers as of now have spam sifting set up, which will move suspicious email to a different garbage organizer. By revealing, blocking and erasing occurrences of spam email that do make it into their inboxes, clients can prepare the customer to keep additionally messages from those specific spam addresses or messages showing comparative substance as shown in figure 2.2.

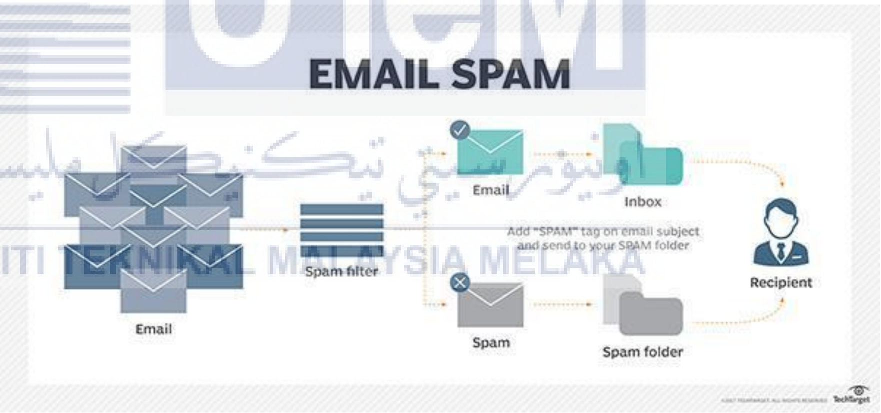


Figure 2.2 : Example of emailspam

For additional assurance, clients can likewise include an outsider hostile to spam channel on neighborhood email customers or make an email whitelist, which incorporates the majority of the particular email addresses, IP locations or spaces the client trusts and will get email from. The white rundown must be completely and persistently refreshed, and it can be a tedious and troublesome process.

Clients who need to distribute their email addresses on the web, for example, in online gatherings or remarks areas, ought to utilize a disposable email account or covered email address.

2.2.2 Statistic about Cyber security

Cyber security or data innovation security are the methods of ensuring PCs, systems, projects and information from unapproved get to or assaults that are gone for abuse.

Picture spam quickly wind up noticeably known as an intend to effectively dodge any printed analyzer implanting in the distinctive spam channels (Zhen et al. 2009; Kelly 2007). The procedure of including a picture rather than content in spam messages begun in 2004 (Kelly 2007) so that up to late 2005, 1% of all spam messages were picture spams (Soranamageswari and Meena 2010). In 2006 and 2007 an awesome development occured so that 27% and 65% of all spam messages are accounted for as picture spam, individually (Mehta et al. 2008; Gao et al. 2008). This upsurge implies that picture spam solidly settled its sweeping statement among cybercrime.

Since the analysts have attempted a great deal for identifying picture spam messages among the true blue messages, the volume of picture spam diminished in 2008 and 2009, generally to 40% of all spam messages (Zuo et al. 2009). In spite of the fact that, due to the new spam traps, picture spammers won the fight again and picture spams are again on the ascent, so that the picture spam messages as announced in SYmantec (Antivi-rus, Anti-Spyware, accessible at http://www.symantec.com) are 55% of spam messages in 2010.

These insights urge the scientists to focus on picture spam more than before and it will be more extensive by realizing that the spam messages are 85% of all messages.

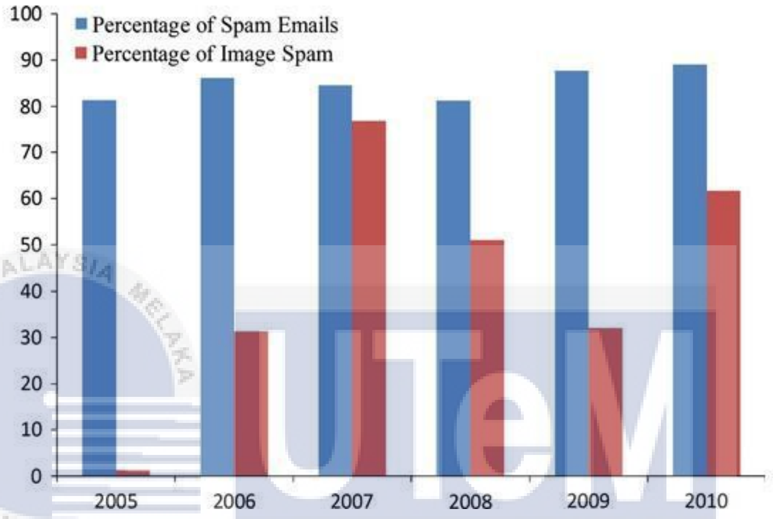


Figure 2.3 : Example graph of Statistic

Based on figure 2.3, In 2010 (Secure Web Gateway—Internet Security and Email Security Solutions, accessible at http://www.m86security.com). Figure 1 demonstrates the rate of picture spams in the greater part of the traded messages which is accounted for by SYmantec (Antivirus, Anti-Spyware, accessible at http://www.symantec.com).

2.2.3 How to prevent Spam?

2.2.3.1 Machine Learning

Machine learning is a kind of manmade brainpower (AI) that gives PCs the capacity to learn without being unequivocally modified. Machine learning centers around the advancement of PC projects that can change when presented to new information.

The procedure of machine learning is like that of data mining. Both frameworks scan through information to search for examples. In any case, rather than extricating information for human appreciation just like the case in information mining applications.

Machine learning utilizes that information to detect designs in information and modify program activities as needs be. Machine learning calculations are frequently ordered as being administered or unsupervised. Supervised algorithms can apply what has been realized in the past to new data. Unsupervised calculations can draw surmisings from datasets.

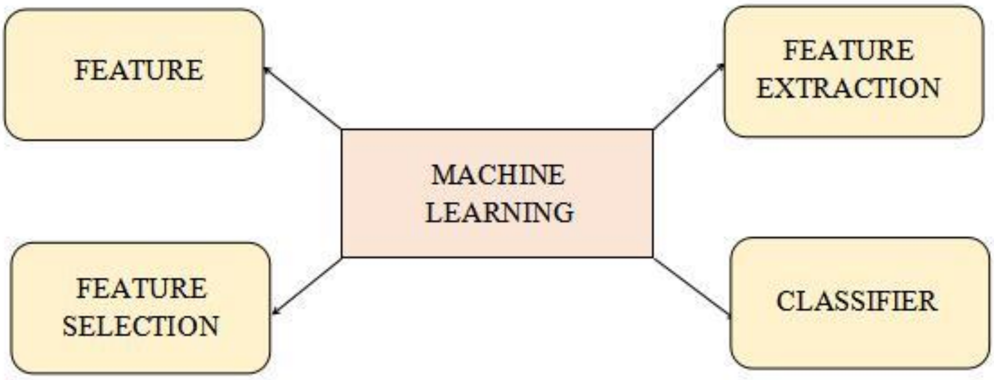


Figure 2.4 : Machine Learning attribute

A. Features

A feature is an individual quantifiable property of a marvel being watched. Picking instructive, segregating and independent features is an essential stride for compelling calculations in example acknowledgment, order and relapse.

A quality is a useful element. An element is only a trademark or property. It is not really useful. An element of my home is its irregular shading. A trait of my home is it has a home theater. It truly doesn't, only a case.

Grouping is the task of an arrangement of perceptions into subsets (called bunches) so that perceptions in a similar bunch are comparative in some sense. Grouping is a strategy for unsupervised learning, and a typical procedure for factual information investigation utilized as a part of many fields.

B. Feature Selection

Include choice is likewise called variable determination or property choice. It is the programmed choice of qualities in your information, (for example, segments in forbidden information) that are most applicable to the prescient demonstrating issue you are dealing with.

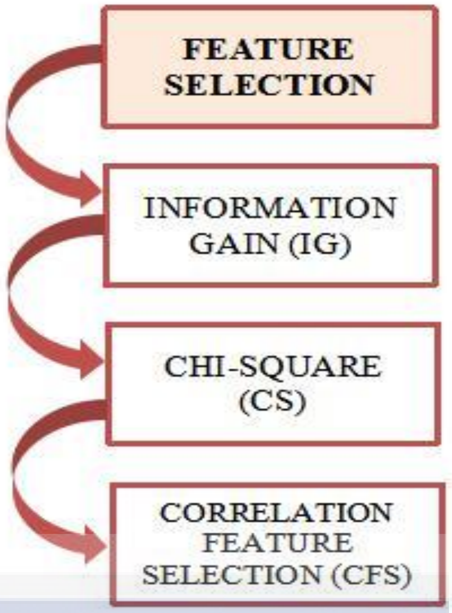


Figure 2.5: Example of Feature Selection

I. Information Gain (IG)

Data gain is an equivalent word for Kullback–Leibler disparity. Be that as it may, with regards to choice trees, the term is in some cases utilized synonymously with mutual data, which is the normal estimation of the Kullback–Leibler difference of the univariate likelihood dispersion of one variable from the restrictive conveyance of this variable given the other one.

II. Chi-Square (CS)

It is a will be a measurable test connected to the gatherings of unmitigated components to assess the probability of relationship or relationship between them utilizing their recurrence dissemination.

III. Correlation Feature Selection (CFS)

CFS (Correlation based Feature Selection) is a calculation that couples this assessment equation with a proper relationship measure and a heuristic hunt procedure. CFS was assessed by investigations on fake and characteristic datasets. CFS commonly wiped out well over a large portion of the elements. CFS executes commonly speedier than the wrapper, which enables it to scale to bigger datasets.

C. Feature Extraction

Include extraction starts from a fundamental course of action of measured data and manufactures decided qualities (highlights) proposed to illuminate and non-monotonous, empowering the subsequent learning and hypothesis steps, and now and again provoking better human interpretations. Include extraction is related to dimensionality diminishment.

Exactly when the data to a count is excessively broad, making it impossible to ever be readied and it is suspected to be dull (e.g. a comparative estimation in both feet and meters, or the monotony of pictures shown as pixels), then it can be changed into a decreased course of action of elements (moreover named an element vector). Choosing a subset of the hidden components is called highlight assurance. The picked elements are required to contain the relevant information from the data.

D. Classifier

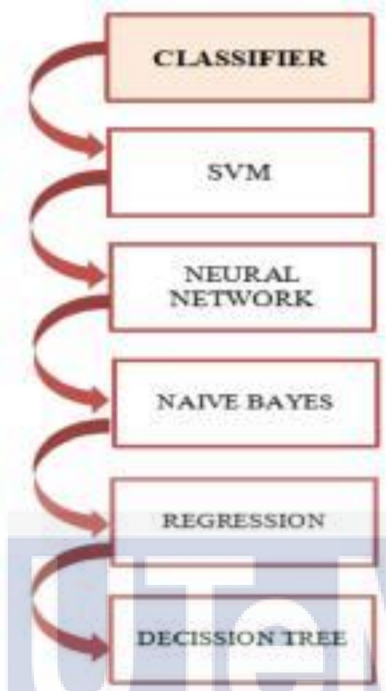


Figure 2.5 : Example of Classifier

I. Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally characterized by an isolating hyperplane. At the end of the day, given named preparing information (regulated taking in), the calculation yields an ideal hyperplane which sorts new cases.

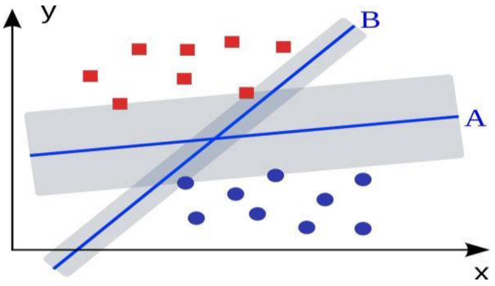


Figure 2.6: Example of SVM

II. Neural Network

A neural system is an arrangement of equipment as well as programming designed after the operation of neurons in the human mind. Neural systems likewise called fake neural systems are an assortment of deep learning technologies. Business uses of these advances by and large concentrate on taking care of complex signal processing or design acknowledgment issues.

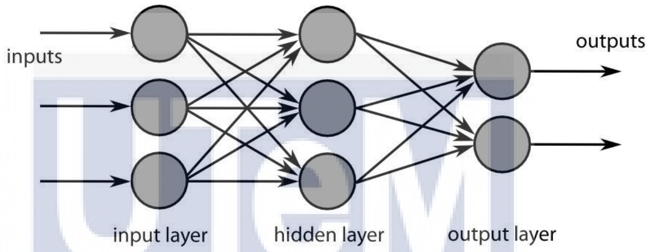


Figure 2.7: Example of Neural Network

III. Naive Bayes

Naive Bayes classifiers are a group of straightforward probabilistic classifiers in light of applying Bayes' hypothesis with solid (guileless) freedom suppositions between the elements.

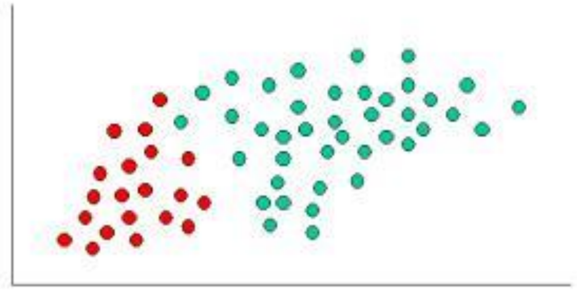


Figure 2.8: Example of Naive Bayes

IV. Regression

Regression analysis is a measurable procedure for assessing the connections among factors. It incorporates numerous procedures for demonstrating and investigating a few factors, when the emphasis is on the connection between a reliant variable and at least one free factors (or 'indicators').

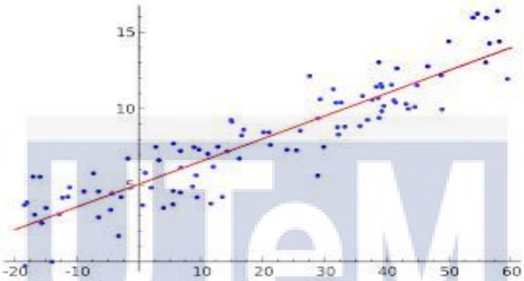


Figure 2.9: Example of Regression

V. Decission Tree

A decision tree is a decision support instrument that uses a tree-like diagram or model of decisions and their conceivable results, including chance occasion results, asset expenses, and utility. It is one approach to show a calculation.

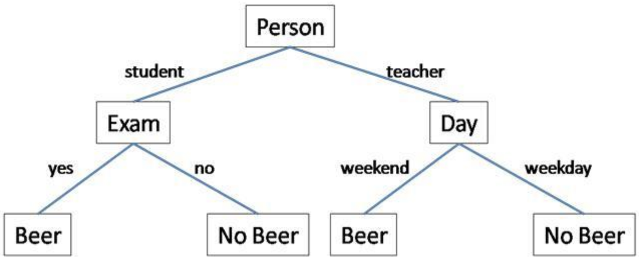


Figure 2.10: Example of Decission Tree

VI. Process Prediction

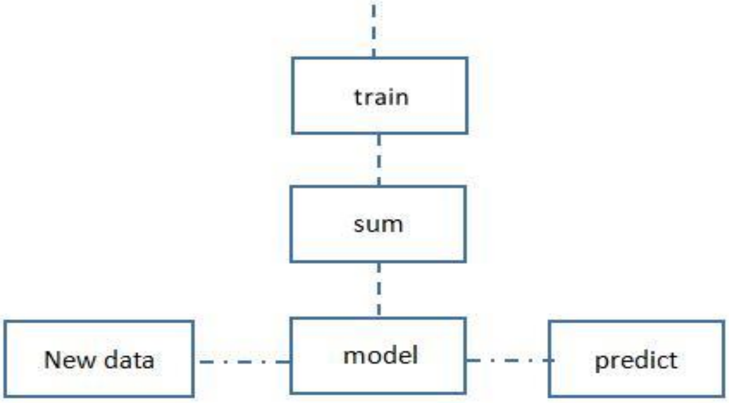


Figure 2.11: Process of prediction

2.2.3.2 White list

A white list is a list of email locations or area names from which an email blocking project will enable messages to be gotten. Email blocking programs, likewise called a spam channels, are planned to avert most spontaneous email messages (spam) from showing up in endorser inboxes. In any case, these projects are not great. Shrewdly created spam gets past, and a couple fancied messages are blocked. Most Internet clients can endure the periodic spontaneous email promotion that a spam channel misses, yet are worried by the prospect that a critical message won't not be gotten. The white white list choice is an answer for the last issue. The white list can be progressively assembled over a timeframe, and can be altered at whatever point the client needs.

Some spam channels erase suspected garbage email messages straightaway, however others enable the client to place them in an isolated inbox. Occasionally, the isolated messages are seen to check whether any of them are real messages. This choice is utilized by some Web-based email customers set up of, or notwithstanding, a white white list.

2.2.3.3 Black list

A black list, in some cases basically alluded to as a boycott, is the production of a gathering of ISP locations known to be sources of spam, a kind of email all the more formally known as spontaneous business email (UCE). The objective of a dark gap rundown is to give a rundown of IP addresses that a system can utilize to filter out undesirable movement. Subsequent to separating, activity coming or heading off to an IP address on the rundown just vanishes, as though it were gulped by an astronomical black opening. The Mail Abuse Prevention System (MAPS) Real-time Black opening List (RBL), which has more than 3000 passages, is a standout amongst the most well known dark gap records. Started as an individual venture by Paul Vixie, it utilized by many servers around the globe. Other prominent dark gap records incorporate the Relay Spam Stopper and the Dial up User List.

2.2.4 Image

In this research,we are concentrate on picture. Picture have a particular elements which is Sift Keypoint. A Sift keypoint is a round picture district with an introduction. It is depicted by a geometric frame of four parameters: the keypoint focus coordinates x and y, its scale (the span of the area), and its orientation (an point communicated in radians). The filter locator utilizes as keypoints picture structures which take after "blobs". Via looking for blobs at numerous scales and positions, the filter indicator is invariant (or, all the more precisely, covariant) to interpretation, pivots, and re scaling of the picture.

The keypoint introduction is likewise decided from the nearby picture appearance and is covariant to picture revolutions. Contingent upon the symmetry of the keypoint appearance, deciding the introduction can be vague. For this situation, the filter identifiers gives back a rundown of up to four conceivable introductions, building up to four edges (varying just by their introduction) for each recognized picture blob.

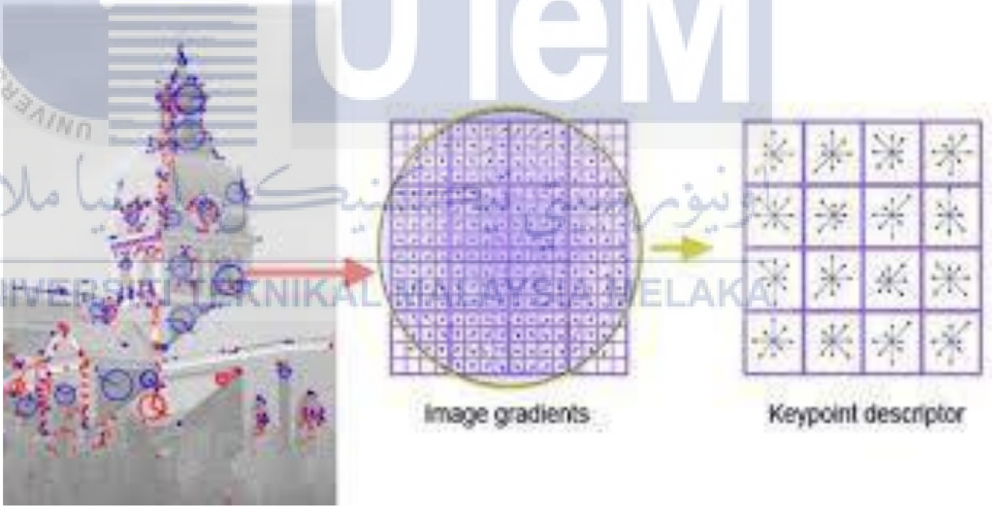


Figure 2.12 : Example of SIFT Keypoint

2.2.5 Method

A. Frequent Itemset Mining

Frequent Itemset Mining is a branch of information mining procedures. The basic thought of FIM is to find fascinating relations between things in extensive database. FIM was initially utilized as a part of the market bushel investigation where an exchange information recorded by a general store is utilized to distinguish sets of items that have been bought.

Utilizing these data, mining affiliation rules find regularities between items. For instance, an affiliation manage {bread, biscuits}=> {margarine}, implying that a client tends to purchase margarine on the off chance that she or he purchases breads and scones together. This data is helpful in deciding basic leadership for showcasing exercises, for example, item positions and limited time cost. Along these lines affiliation rules mining can be separated into two stages, in the first place, mining frequents itemset to produce all substantial affiliation rules.

FIM has found to extensively utilized as a part of uses in regions, for example, web utilization mining., interruption recognition and bioinformatics. Despite the fact that FIM produces sets of discriminative components, shockingly, it is not much of the time utilized as a part of picture arrangement strategies. In this paper we don't utilize FIM to produce affiliation rules. Rather we need to get an arrangement of all successive itemset showing up at any rate with a base bolster limit in the dataset. At that point, these successive itemset will utilized as highlight descriptors to depict pictures and apply them to the learning calculation.

B. Ensemble Method

Ensemble methods are learning algorithms that construct a. set of classifiers and then classify new data points by taking a (weighted) vote of their predictions. The original ensemble method is Bayesian aver- aging, but more recent algorithms include error-correcting output coding, Bagging, and boosting.

C. K-mean Clustering

K-mean clustering is a sort of unsupervised realizing, which is utilized when you have unlabeled information (i.e., information without characterized classifications or gatherings). The objective of this calculation is to discover bunches in the information, with the quantity of gatherings spoken to by the variable K. The calculation works alliterative to allocate every information indicate one of K groups in light of the components that are given. Information focuses are grouped in light of highlight similitude. The aftereffects of the K-means bunching calculation are the centroids of the K clusters, which can be utilized to mark new information and names for the preparation information (every information indicate is alloted a solitary group).

As opposed to characterizing bunches before taking a gander at the information, grouping enables you to discover and examine the gatherings that have framed naturally. The "picking K" segment beneath portrays how the quantity of gatherings can be resolved. Every centroid of a bunch is an accumulation of highlight qualities which characterize the subsequent gatherings. Inspecting the centroid include weights can be utilized to subjectively decipher what sort of gathering each bunch speaks to.

K-mean clusteringcalculation covers a typical business cases where K-means is utilized, the means required in running the calculation and a python illustration utilizing conveyance armada information.

2.3 Previous Researcher

2.3.1 A survey of learning-based techniques email spam filtering

Email spam is one of the significant issues of the today's Internet, conveying finan-cial harm to organizations and irritating individual clients. Among the methodologies created to stop spam, sifting is an imperative and mainstream one. In this paper we give a diagram of the cutting edge of machine learning applications for spam sifting, and of the methods for assessment and examination of various separating strategies. We likewise give a short descrip-tion of different branches of hostile to spam security and talk about the utilization of different methodologies in business and non-business against spam programming arrangements.

2.3.2 A survey of image spamming and filtering techniques

Numerous systems have been proposed to battle the upsurge in picture based spam. All the proposed systems have a similar target, attempting to evade the picture spam entering our inboxes. Picture spammers dodge the channel by various traps and each of them should be broke down to figure out what office the channels need for beating the traps and not enabling spammers to full our inbox. Diverse traps offer ascent to various systems. This work reviews picture spam marvels from all sides, containing definitions, picture spam traps, hostile to picture spam procedures, informational collection, and so forth. We depict each picture spamming trap independently, and by examining the strategies utilized by specialists to battle them, an arrangement is attracted three gatherings: header-based, content-based, and message based. At long last, we disk the informational indexes which specialists use in trial assessment of their articles to demonstrate the precision of their thoughts.

2.3.3 A review of machine learning approaches to Spam filtering

In this paper, we exhibit a far reaching survey of late advancements in the use of machine learning calculations to Spam separating, concentrating on both printed and picture based methodologies. Rather than considering Spam sifting as a standard grouping issue, we highlight the significance of consider-ing particular qualities of the issue, particularly idea float, in outlining new channels. Two partic-ularly essential perspectives not broadly perceived in the writing are examined: the troubles in refreshing a classifier in light of the pack of-words portrayal and a noteworthy distinction between two early guileless Bayes models. Generally speaking, we presume that while critical headways have been made in the most recent years, a few angles stay to be investigated, particularly under more sensible assessment settings.

2.3.4 Machine Learning Approaches for Modeling Spammer Behavior

Spam is normally known as spontaneous or undesirable email messages in the Internet making potential risk Internet Security. Clients invest an important measure of energy erasing spam messages. All the more imperatively, steadily expanding spam messages possess server storage room and devour organize transfer speed. Catchphrase based spam email sifting techniques will in the end be less fruitful to model spammer conduct as the spammer continually changes their traps to evade these channels. The sly strategies that the spammer uses are examples and these pat-terns can be demonstrated to battle spam. This paper researches the potential outcomes of displaying spammer behavioral examples by surely understood grouping calculations, for example, Naïve Bayesian classifier (Naïve Bayes), Decision Tree Induction (DTI) and Support Vector Machines (SVMs). Preparatory exploratory outcomes exhibit a promising recognition rate of around 92%, which is significantly an upgrade of execution contrasted with comparable spammer conduct demonstrating research.

2.4 Conclusion

In this chapter, it explained about the type of spam that attack in daily life such as image spam, email spam, web spam, sms spam and social spam. Other than that, it also define on how to prevent the spam by using the filtering techniques. Then, in this chapter will explained on what techniques that will use to detect an image spam. There is also contained the previous research section which in this section it will explained on what previous researcher had been done on their researcher about detecting image spam. The researcher will explained on what techniques that they used to detect an image spam so we can compared their techniques.

CHAPTER III

METHODOLOGY

3.1 Introduction

At first, the spam e-mails are text-based and manipulate various text spam tricks including text splitting. Encoding abuses, attack on tokenizer and symbolic text. In response, many effective text-based anti-spam filters were proposed, resulting in difficulties for the spam e-mails to pass through these filters. Spammers made attempts to outsmart the text based filtering by embedding texts into images.

As the content-based filtering use the image processing techniques, several image features such as colour, edge and texture are usually exploited by the image spam filters. A number of studies have demonstrated that colour features are among the most important factors and provide compact representation of images. Even though the Scale Invariant Feature Transform (SIFT) as proposed is the most widely used image descriptor, there are very few studies that investigated its impact on image spam filtering.

3.2 Project Methodology

In this section, we describe about the purposed method which involves how generate BOW feature vectors, the process of extracting FIM descriptors from BOW, and finally, how we combine FIM classifiers with other classifiers as inputs to the ensemble methods. Figure 1 shows a block diagram on how models for FIM.generated. There are three main stages, namely the vector quantisation, feature selection and classification.

Vector quantisation start by identifying the keypoints in the image based on SIFT algorithm. Once the keypoints have been identified, the keypoints descriptor are created. After that the keypoints dataset are build which involves constructing a keypoint dataset that the k-means clustering algorithm will work on. Since we use a SIFT as a local feature, each SIFT descriptor has a 128 features.

These 128 features form a 128-dimensional feature vector which uniquely represent a keypoint. In this step, a keypoint dataset consisting of all keypoint feature vectors that are extracted from the images is generated. Then the k-means clustering algorithm is applied to the keypoint feature vectors.

Clustering tends to group more similar SIFT descriptors within the same cluster. The k-mean algorithm takes the feature vectors and the number of clusters to generate, k, as the input and return a set of cebtroids. These centroids have the same feature dimension as the keypoint feature vectors. A codebook mapping the cluster numbers and centroid is generated in this stage.

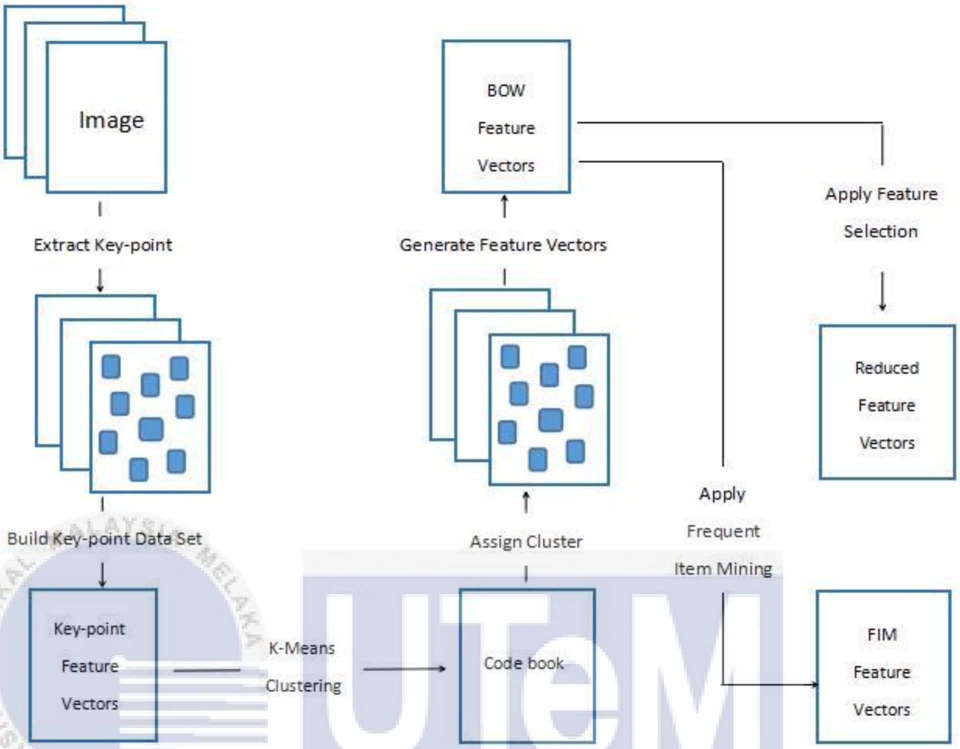


Figure 3.1: Block diagram to generate models

After the codebook is generated, the distance between a keypoint and centroids are computed. The keypoint as assigned to a centroid to which it is the closet. This assignment is based on the minimum sum of the squareddistances between a keypoint and the centroids. However, to simplify the representation, each keypoint is represented by a cluster number rather than its centroid.

In the feature selection stage, three different approaches are applied which is first, to classify BOW feature vectors without going through any feature selection processes. Second, BOW feature vectors will be applied. Third, frequent itemsets in the BOW feature vectors are identified and a new FIM feature vector will be generated.

The resulting feature vectors, after going through the feature selection process using the same feature descriptors as BOW but the size has been reduced. The selected feature descriptors depends on the feature selection algorithm.However, the new feature vectors are applied to the frequent item mining.

We use a small example to illustrate how the FIM feature descriptors are extracted from the BOW feature vectors. Refer to table 3.1 , the set of features are cluster 2 (C1), cluster 2 (C2), cluster 3 (C3), cluster 4 (C4), cluster 5 (C5) while the number of image are 5.

3.2.1 Scale Invariant Feature Transform (SIFT) Keypoint

SIFT is an algorithm developed by David Lowe in 2004 for the extraction of interest points from gray-level images.

The input is a gray-level image. The output is a list of 2D points on the image each associated to a vector of low-level descriptors. These points are said keypoints and their descriptors are invariant by rescaling, in-plane rotating, noise addition and in some cases by changes of illuminant.

Keypoints provide a local image description. They are used to find visual correspondences between images for different applications, like image alignment or object recognition.

Since we used SIFT as a local feature, each SIFT descriptor has 128 features. These 128 features form a 128-dimensional feature vectors which uniquely represents a keypoint.

The figure 3.2 shows the explanation that focus on SIFT keypoint.

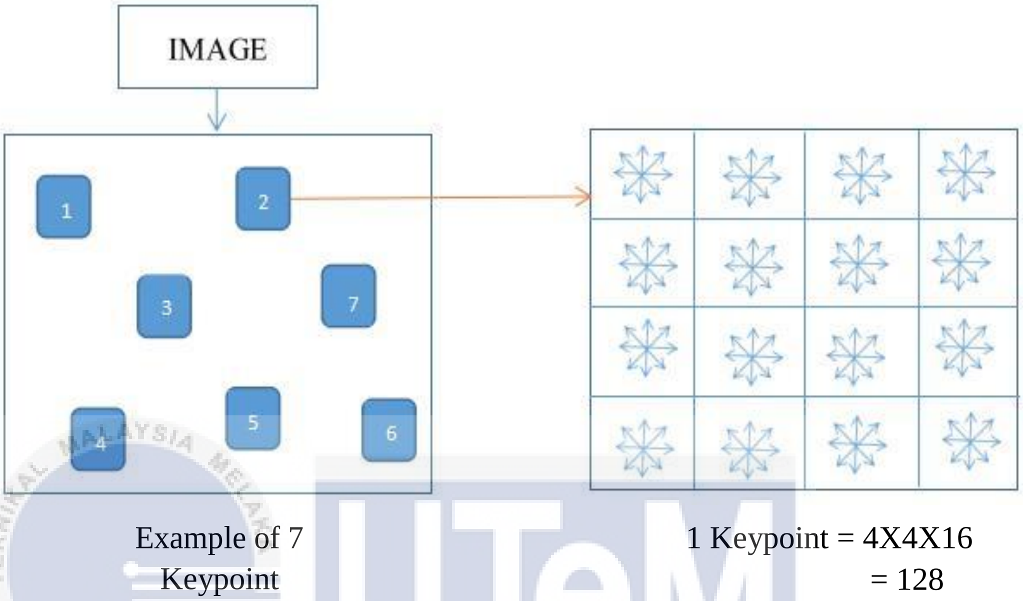


Figure 3.2 : Example of keypoint

The figure 3.3 illustrate the keypoint after make a clustering.

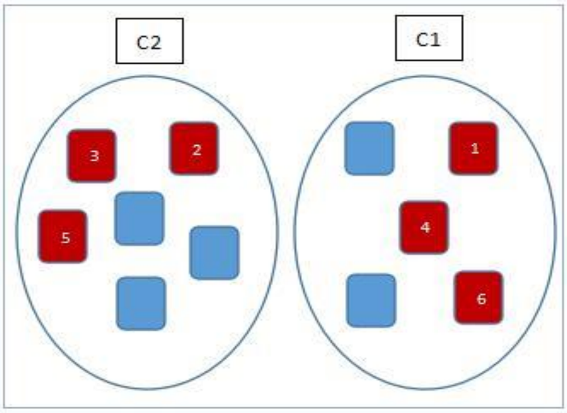


Figure 3.3 : Example of keypoint

3.2.2 Scale Invariant Feature Transform (SIFT) Descriptor

This descriptor as well as related image descriptors are used for a large number of purposes in computer vision related to point matching between different views of a 3-D scene and view-based object recognition. The SIFT descriptor is invariant to translations, rotations and scaling transformations in the image domain and robust to moderate perspective transformations and illumination variations. Experimentally, the SIFT descriptor has been proven to be very useful in practice for image matching and object recognition under real-world conditions.

In its original formulation, the SIFT descriptor comprised a method for detecting interest points from a grey-level image at which statistics of local gradient directions of image intensities were accumulated to give a summarizing description of the local image structures in a local neighbourhood around each interest point, with the intention that this descriptor should be used for matching corresponding interest points between different images. Later, the SIFT descriptor has also been applied at dense grids (dense SIFT) which have been shown to lead to better performance for tasks such as object categorization, texture classification, image alignment and biometrics . The SIFT descriptor has also been extended from grey-level to colour images and from 2-D spatial images to 2+1-D spatio-temporal video.

3.3 Frequent Item Mining (Previous Researcher)

In previous researcher, his generate a list of 2-pairs of the frequent items. The candidate of 2-pair frequent items are only selected from a pool of frequent items which consists of the sets {C2-C3}, {C3-C4}, and {C4-C5}. Previous researcher used the MinimumWeightage method to make a clustering based on Table 3.1.

3.3.1 BOW Feature Vector and Generated from BOW (Minimum Weightage)

Table 3.1 : Example of BOW Feature Vector

Image Number	C1	C2	C3	C4	C5
1	0	1	2	1	1
2	3	1	2	0	0
3	0	0	0	2	1
4	2	1	2	2	1
5	0	0	2	2	0

Minimum Weightage 1

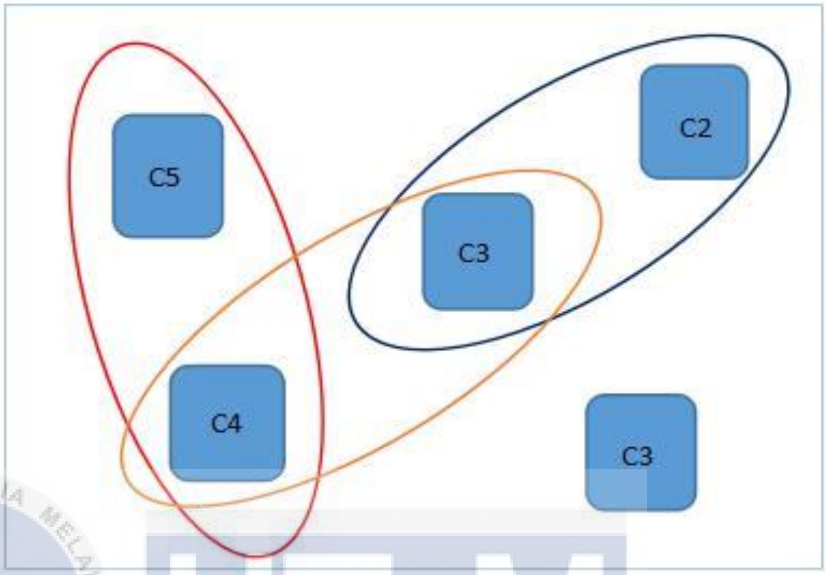


Figure 3.4: Image for Minimum Weightage 1

Table 3.2: Frequncy from Minimum Weightage in image 1

Image 1	C2-C3	C3-C4	C4-C5
Frequency	1	1	1

Minimum Weightage 2

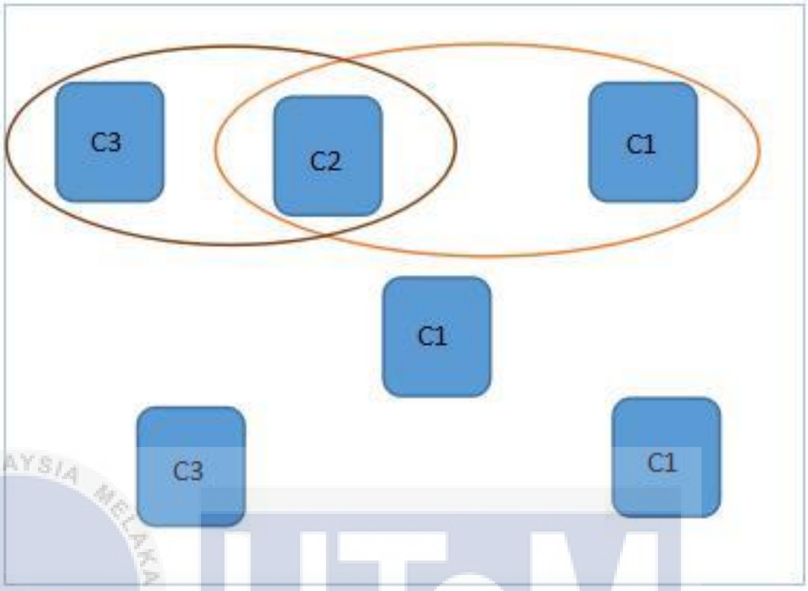


Figure 3.5: Image for Minimum Weightage 2

Table 3.3 : Frequency from Minimum Weightage in Image 2

Image 2	C1-C2	C2-C3
Frequency	1	1

Minimum Weightage 3

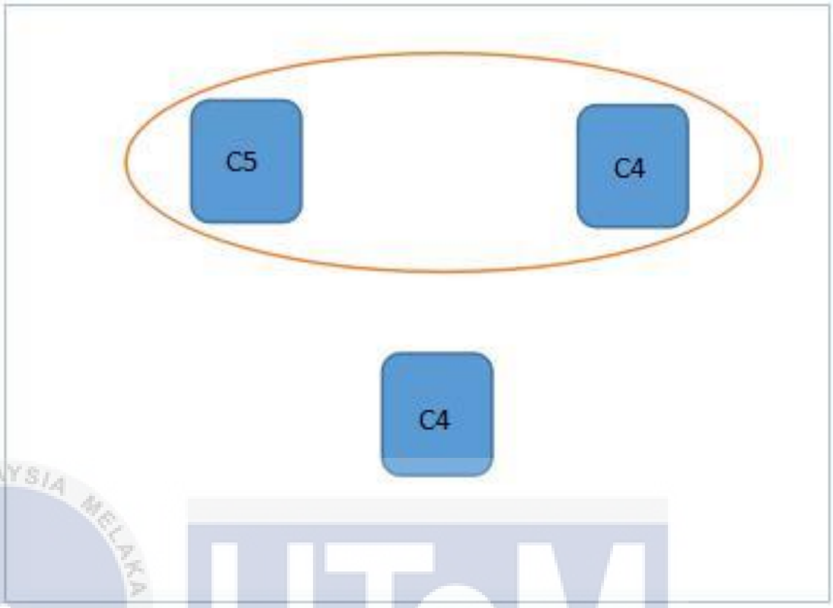


Figure 3.6: Image for Minimum Weightage 3

Table 3.4 : Frequency from Minimum Weightage in Image 3

Image 3	C4-C5
Frequency	1

Minimum Weightage 4

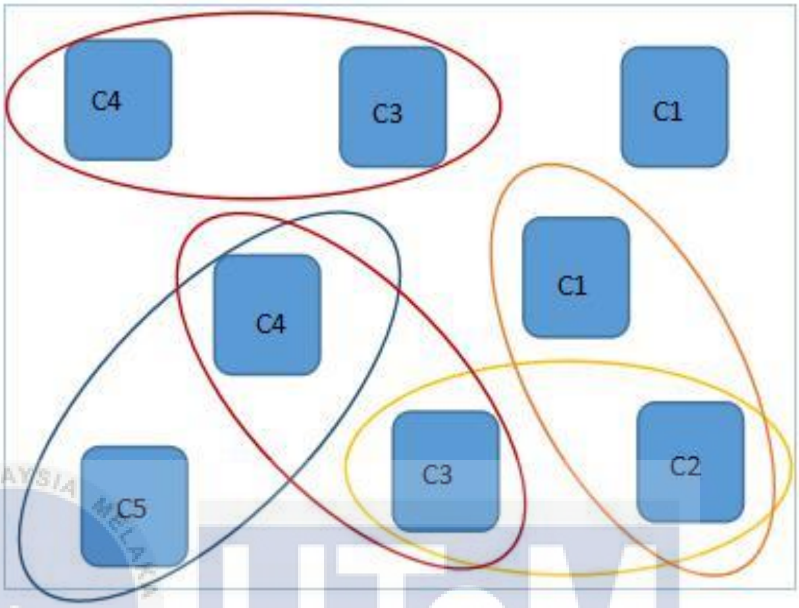


Figure 3.7: Image for Minimum Weightage 4

Table 3.5 : Frequency from Minimum Weightage in Image 4

Image 4	C1-C2	C2-C3	C3-C4	C4-C5
Frequency	1	1	2	1

Minimum Weightage 5

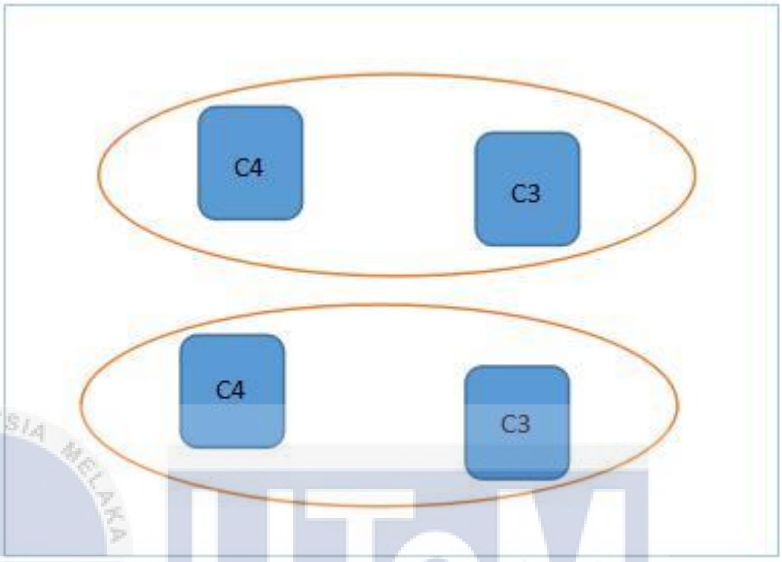


Figure 3.8: Image for Minimum Weightage 5

Table 3.6 : Frequency from Minimum Weightage in Image 5Image 5

Table 3.3 :	C3-C4
Frequency	2

3.4 Frequent Item Mining (New Research)

In this research, will generate a list of 2-pairs or more of the frequent items based on Table 3.7. The candidate of 2-pair or above frequent items will selected from a pool of frequent items which consists of the sets {C2-C3}, {C3-C4}, and {C4-C5}. In this time method used to generate a clustering are Maximum Weightage. This method maybe will help SVM classifier better.

3.4.1 BOW Feature Vector and Generated from

BOW A. Maximum Weightage

Table 3.7 : Example of BOW Feature Vector

Image Number	C1	C2	C3	C4	C5
1	0	1	2	1	1
2	3	1	2	0	0
3	0	0	0	2	1
4	2	1	2	2	1
5	0	0	2	2	0

Maximum Weightage 1

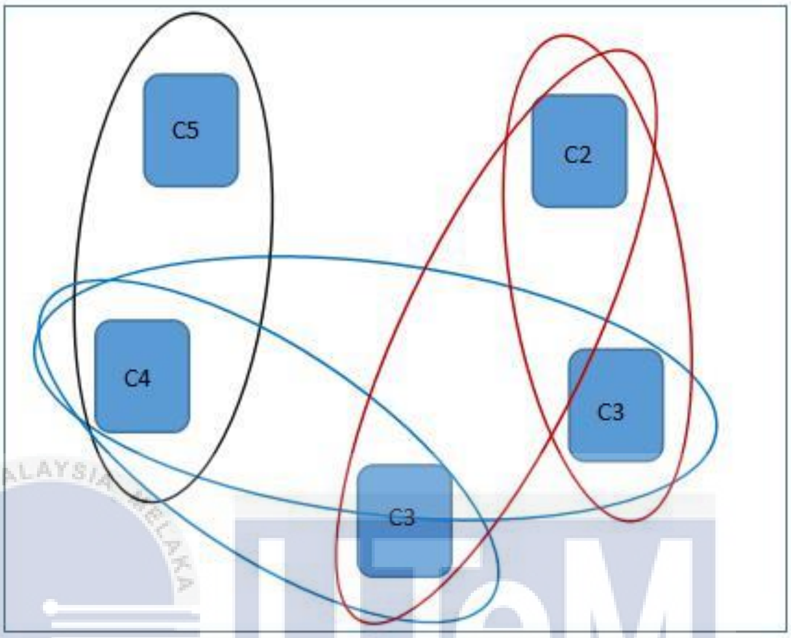


Figure 3.10: Image for Maximum Weightage 1

Table 3.8 : Frequency from Maximum Weightage in Image 1

Image 1	C2-C3	C3-C4	C4-C5
Frequency	2	2	1

Maximum Weightage 2

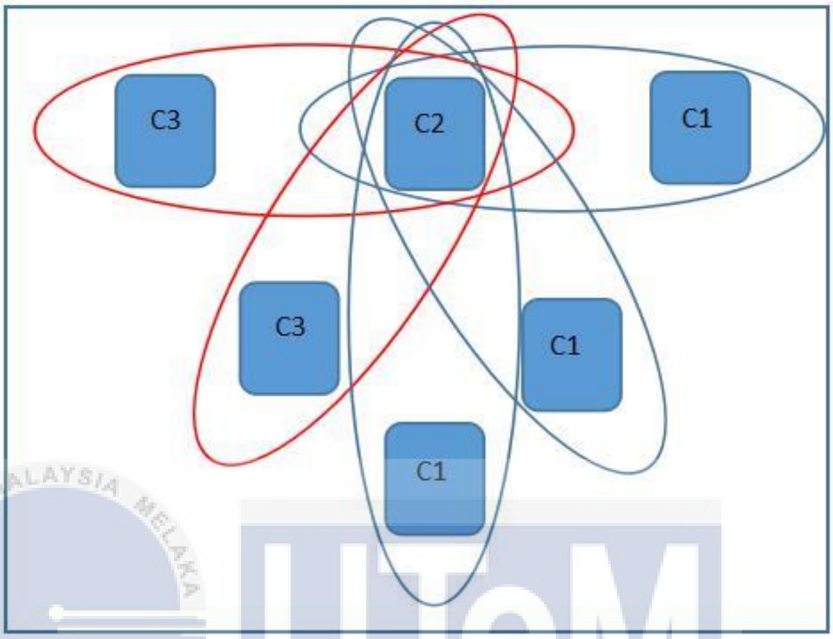


Figure 3.11: Image for Maximum Weightage 2

Table 3.9 : Frequency from Maximum Weightage in Image 2Image

Image 2	C1-C2	C2-C3
Frequency	3	2

Maximum Weightage 3

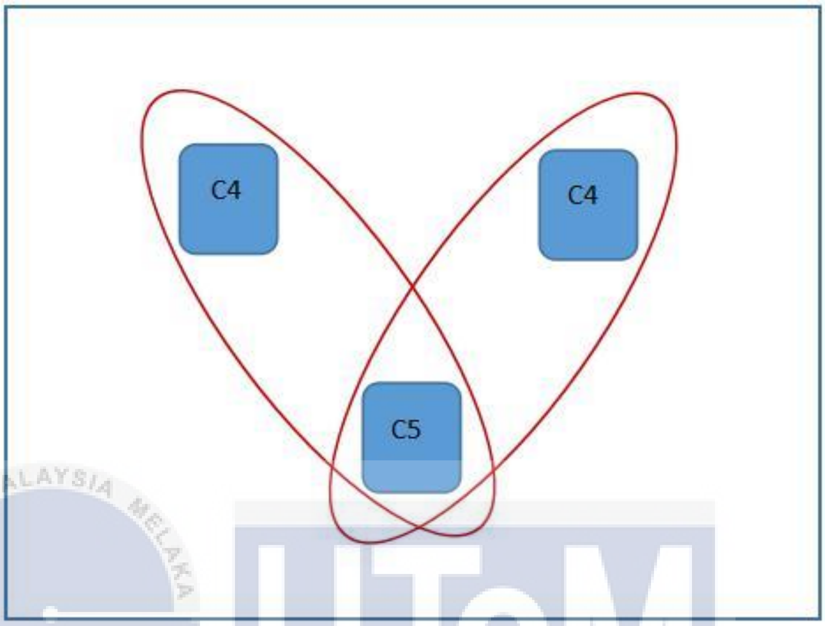


Figure 3.12: Image for Maximum Weightage 3

Table 3.10 : Frequency from Maximum Weightage in Image 3

Image 3	C4-C5
Frequency	2

Maximum Weightage 4

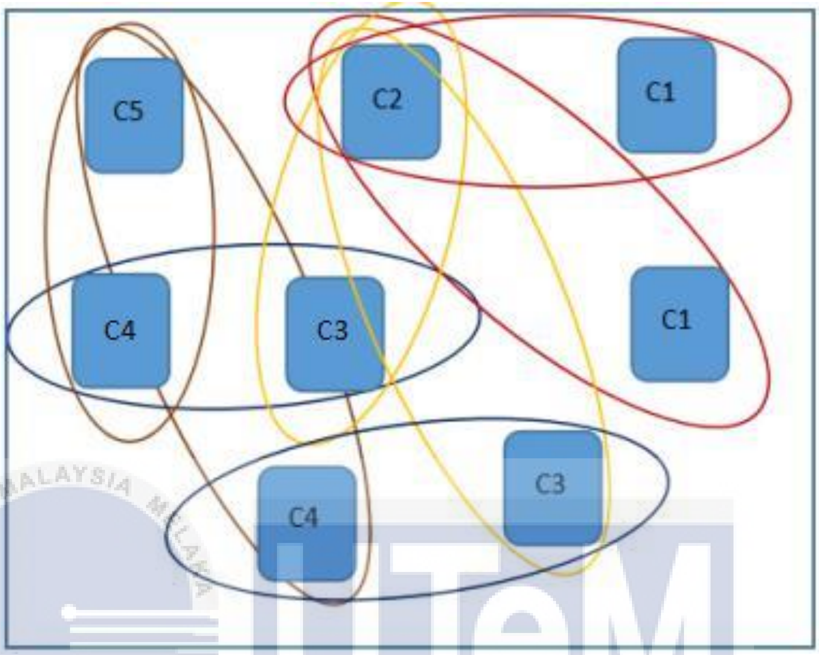


Figure 3.13: Image for Maximum Weightage 4

Table 3.11 : Frequency from Maximum Weightage in Image 4age

Image 4	C1-C2	C2-C3	C3-C4	C4-C5
Frequency	2	2	2	2

Maximum Weightage 5

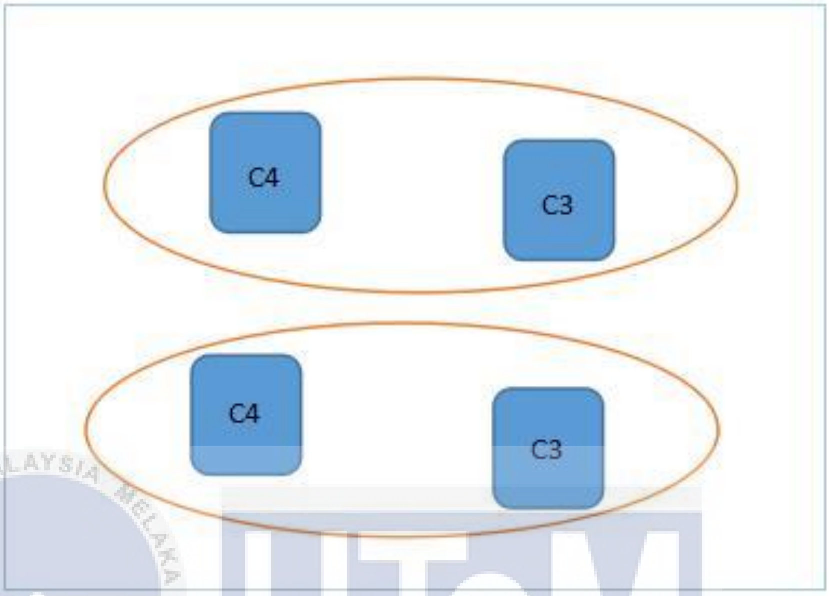


Figure 3.14: Image for Maximum Weightage 5

Table 3.12 : Frequency from Maximum Weightage in Image 5

Image 5	C3-C4
Frequency	2

3.5 Project Schedule and Milestone

3.5.1 Milestone of project

Table 3.13 : Milestone of project

Week	Activity	Action
1	Proposal PSM : submission	Deliverable - Proposal Action - Student
1	Proposal assessment and verification	Action – Supervisor and PSM committee member
2	Proposal correction/ improvement	Student
3	Chapter 1	Deliverable – chapter1 Action – student, supervisor
4	Chapter 2	Action - student
5	Chapter 1 and 2 : submission	Deliverable – Chapter 1 and 2 Action - student
6	Chapter 2 correction / improvement	Action – student
7	Chapter 3 Student Status	Action – PSM committee member, student, supervisor
8	Mid semester break	
9	Project Demo Chapter 3	Action – Student , supervisor
10	Project demo Chapter 3 : submission	Deliverable : chapter 3 Action – Student , supervisor
11	Project demonstration Chapter 4	Action : student

12	Project demo Chapter4 : submission	Deliverable : Chapter 4 Action - Student, supervisor
13	Project demo & presentation report PSM schedule	Deliverable - PSM report Action - Student, supervisor Action – PSM committee member
14	FINAL PRESENTATION	Action – Student, supervisor, evaluator

3.5.2 Gantt Chart

Gantt chart is one of the most useful tools that specialized bar chart useful approach to provide a graphical overview and schedule of all tasks or to indicate the work elements and dependencies of project by showing the activities displayed against time..For this project, Gantt chart is being used to portray the process of the project from the earliest starting point to ending point that fully managed and done completely according to the duration of the time.

Each activity had been scheduled on of each process need to be finished with a specific end goal in order before proceed to the next phase until the last of the activity for this project . The main purpose of Gantt chart is to guarantee this project run smoothly from the earlier starting point . The Gantt chart has been construct in figure below :

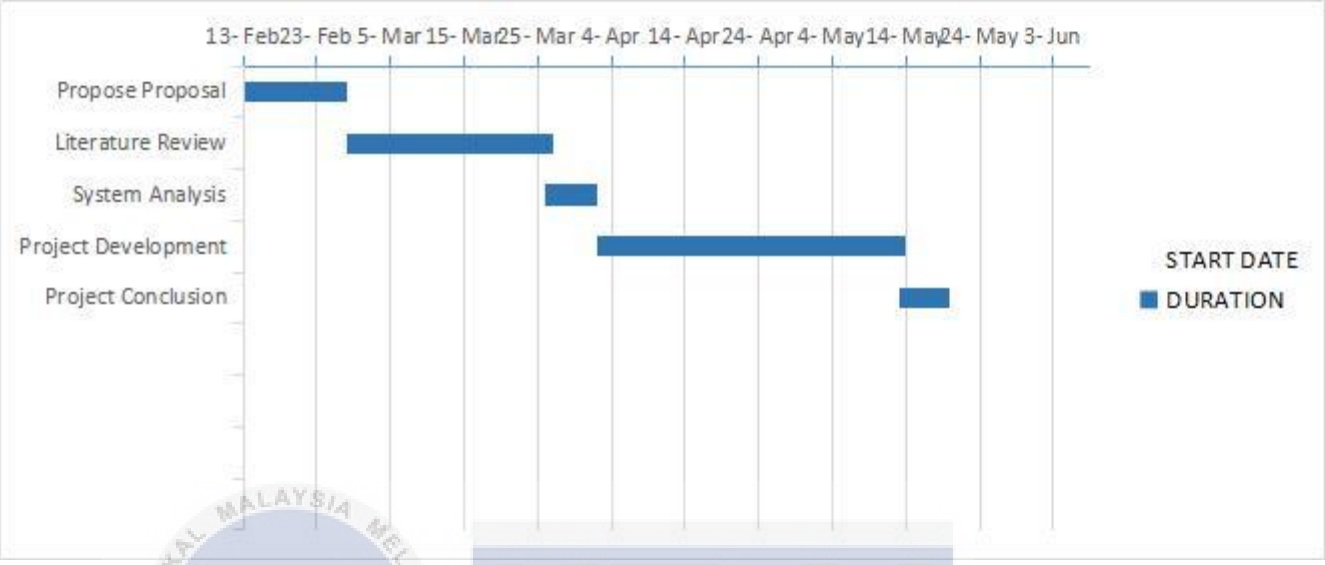


Figure 3.15 Gantt Chart

3.6 Conclusion

The conclusion of this chapter, the methodology describes the steps and processes involved in conducting this project from beginning till the completion which assist in delivering the outcome expected in this project. Thus proper methodology is crucial and it ensures that the project is executed in a systematic way. The subsequent chapter will discuss in detail on the analysis, design, implementation and testing phase mentioned above.

CHAPTER IV

ANALYSIS AND DESIGN

4.1 Introduction

Analysis and design is the important stage during development. It helps to define a clear idea on the research. In this chapter, the analysis and design of image spam detection using Frequent Itemset Mining technique is discussed in detail. This chapter also show the flow of the project by using diagram. The flow are visualized using appropriate diagram. In this chapter are included flow chart, use case diagram and sequence diagram.

4.2 Problem Analysis

The problem analysis that has been identified is summarized in table 4.1 below

Table 4.1 Summaries of Problem Analysis

PS	Problem Analysis
PS1	The effect of image spam

PS1: The effect of image spam.

The examples of the problem is bandwidth used for free and the email users will waste time to delete spam emails or the image spam will attach for virus and also malware.

4.3 Flow Chart

4.3.1 Concept of Flow Chart Diagram

A flowchart is a type of diagram that represents an algorithm, workf low or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows. This diagrammatic representation illustrates a solution model to a given problem.

In this chapter, the flow of classify Frequent Itemset Mining (FIM) will illustrate through a flow chart diagram because flow chart diagram explains the process would be applicable.

4.3.2 Flow Chart Diagram

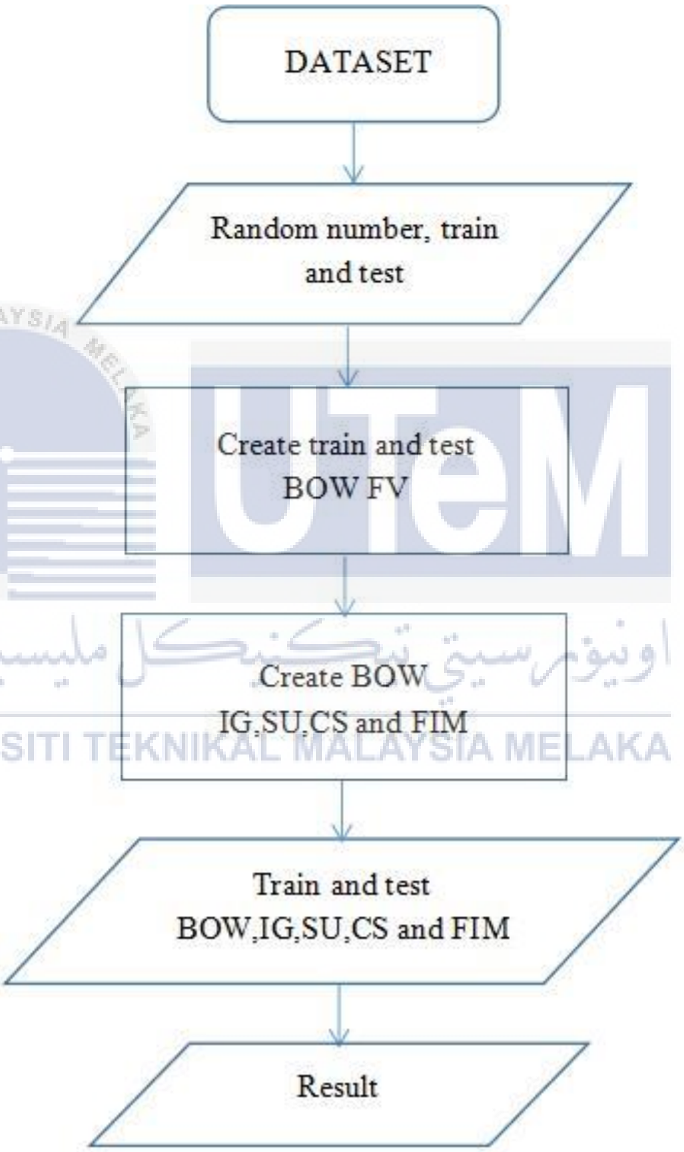

DATASET
Random number, train and test
Create train and test BOW FV
Create BOW IG,SU,CS and FIM
Train and test BOW,IG,SU,CS and FIM
Result

Figure 4.1 : Flow Chart Diagram

4.4 Use Case Model

A use case diagram is a graphic depiction of the interactions among the elements of a system. A use case is a methodology used in system analysis to identify, clarify, and organize system requirements. In this project, there are four things in a data mining module and one thing in a classifier. The use case for data mining module is :

i. Get the dataset.

ii. Pre-processing data.

iii. Training the data in the machine learning.

iv. Read the predict file

4.4.1 Use Case Diagram for Data Mining Module

Use case diagram is a graphic model which display the relation of the system and the user. Use case diagram will construct directly based on the activity stated before. Figure below show the use case diagram for data mining module. There are four use case diagram involved which is get process dataset, pre-processing data, train the machine learning and read .predict.

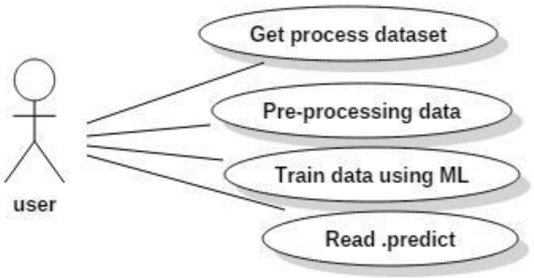


Figure 4.2 : Use Case Diagram for data mining module

Table below shows the narrative of the use case for whole module. The narrative will briefly explain the use action and system action.

Table 4.2 Narrative Use Case for Get Process Dataset

Use case	: Get process dataset.
Author	: User
Purpose	: Process data for the training dataset
Description	: The use case involved 2 process which is read and categorize the data
Before	: Get process dataset
After	: Data is classify

User Action	System Action
1) Key in the dataset into the system	
	1) System will read the list of data 2) System classify the data

Table below shows the narrative of the use case for whole module. The narrative will briefly explain the use action and system action.

Table 4.3 Narrative Use Case for Pre-processing Data

Use Case : Pre-processing data

Actor : User

Purpose : Make a pre-processing data

Description : The use case involved 3 process which is classify the data into spam or ham. This step will repeat until all the data will generate the feature vector. Last step, the result with appear either the data is spam or ham.

Before : Dataset

After : Dataset with frequency for each attribute of spam or ham

User Action	System Action
1) Key in the dataset into the system.	1) System will read the dataset by the cluster number. 2) System classify the dataset either its spam or ham. 3) System create a feature descriptor and give the frequency of feature vectors.

Table below shows the narrative of the use case for whole module. The narrative will briefly explain the use action and system action.

Table 4.4 Narrative Use Case for Train Machine Learning

Use Case : Train machine learning

Actor : User

Purpose : To ensure machine learning know the pattern of data.

Description : User will make a training process for the machine learning so that the data will understand the patern of image spam

Before : Best weightage produced

After : Result file of single classifier and file prediction

User Action	System Action
1) Classify the clustering of image	1) System will train the data using feature vector using Frequent Itemset Mining

Table below shows the narrative of the use case for whole module. The narrative will briefly explain the use action and system action.

Table 4.5 Narrative Use Case for Predict

Use Case : Predict file

Actor : User

Purpose : To ensure the accuracy of result

Description : The data that had been trained before will have testing

Before : Training of prediction file

After : Prediction result

User Action	System Action
1) Choose the data for spam or ham with the same total with training data.	1) System will run the prediction file using the combining classifier and get the result.

4.5 Process Design

4.5.1 Sequence Diagram Concept

A sequence diagram is an interaction diagram that indicates how objects work with each other and in what arrange. It is a develop of a message arrangement diagram.

A grouping graph indicates protest cooperation organized in time arrangement. It delineates the items and classes required in the situation and the grouping of messages traded between the articles expected to complete the usefulness of the situation. Arrangement graphs are ordinarily connected with utilize case acknowledge in the Logical View of the framework being worked on. Grouping outlines are once in a while called event diagrams or event situations.

A grouping outline appears, as parallel vertical lines (life savers), diverse procedures or items that live all the while, and, as flat bolts, the messages traded between them, in the request in which they happen. This permits the detail of basic runtime situations in a graphical way.

4.5.2 Sequence Diagram for Image Spam Detection

Sequence diagram in this project consist of two phase . Phase one is to generate the dataset while phase 2 is to generate the frequent itemset mining method. In the phase 1 shows the interaction between raw dataset , system, SIFT-keypoint , K-mean clustering and feature vector. In this phase the raw dataset will process to generate a dataset. While for the phase 2 it is an illustration of sequence diagram of to generate frequent itemset mining (FIM) method. Frequent itemset mining will help the SVM to make a classifier with better

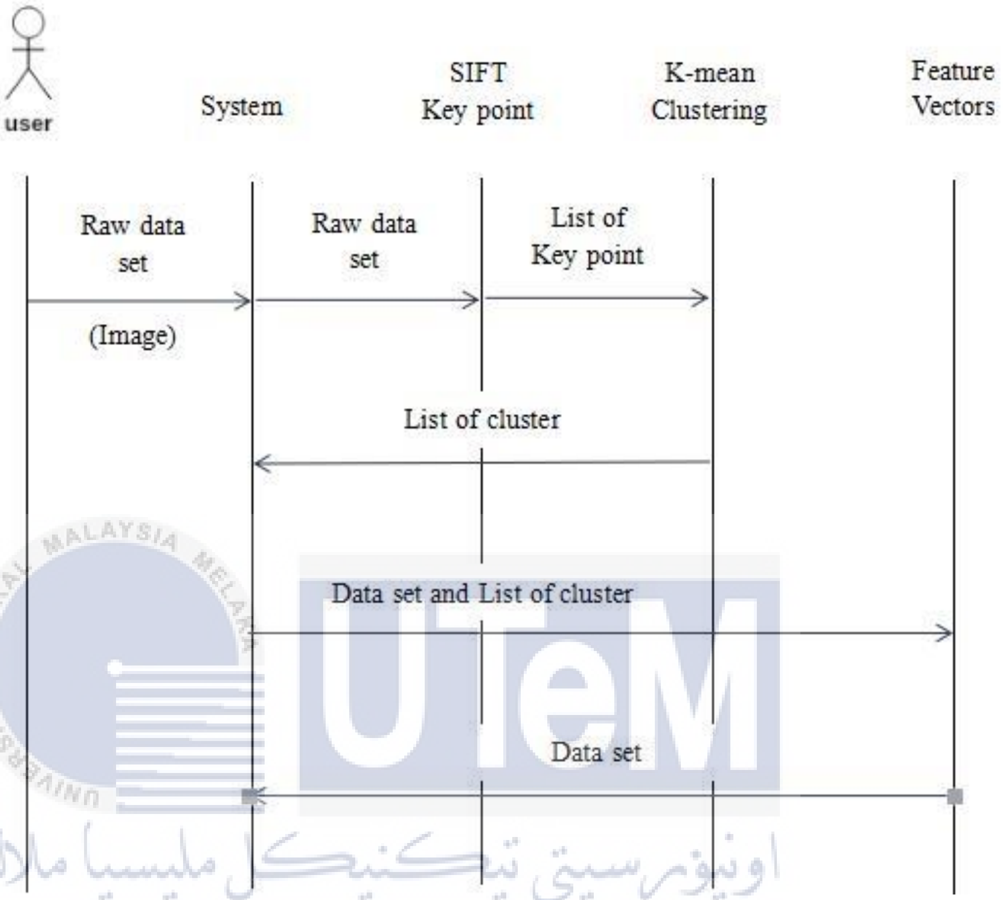


Figure 4.3 : Sequence Diagram phase 1

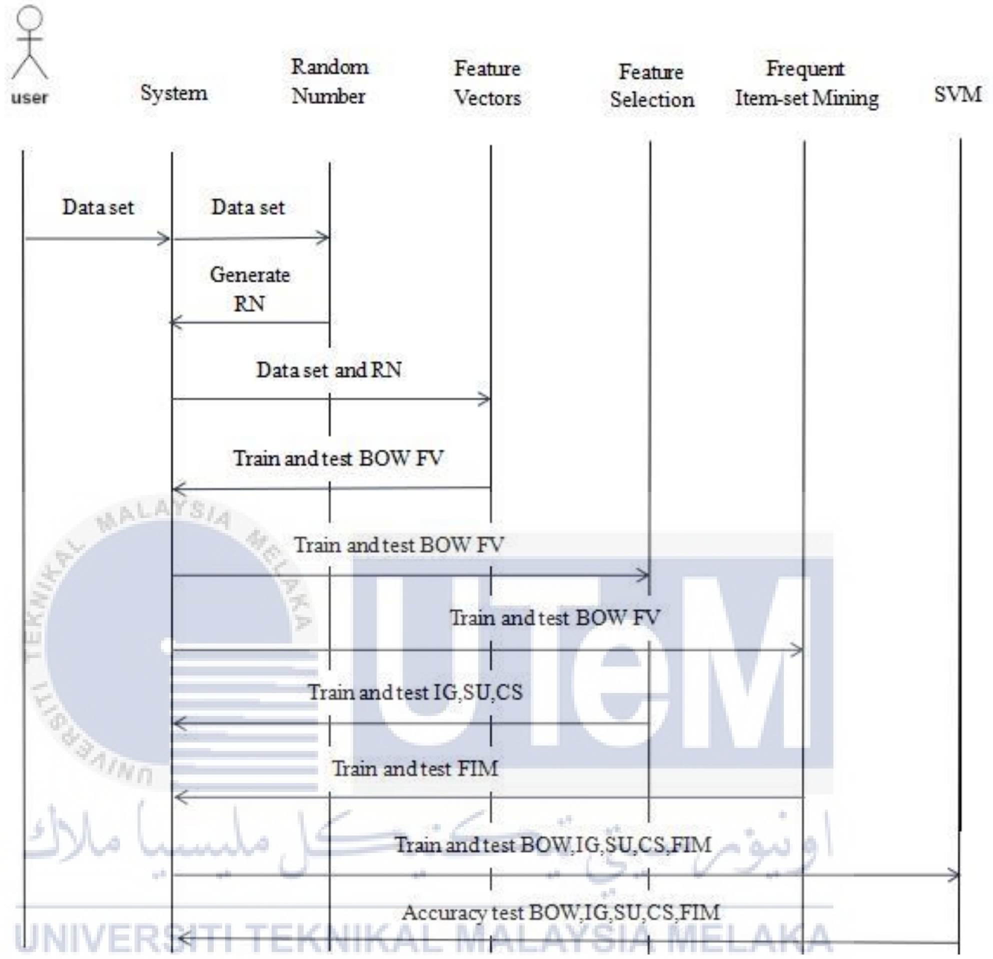


Figure 4.4 : Sequence Diagram phase 2

Legend	
	User
IG	Information Gain
BOW	Bag of word
SU	Symmetrical Uncertainty
CS	Chi-Square
FIM	Frequent Itemset Mining
FV	Feature Vector
RN	Random Number

4.6 Requirement Analysis

For this section, it will analyse the requirement and give all the details about each of the software and hardware that involved in this project. The details will separate by section.

4.6.1 Software Requirement

Table 4.5 Software Requirement

Software	Description
I. Eclipse	Platform to run and test the code and the coding implementation.
II. Windows 10	Workstation for software installment setup.
III. Weka Machine Learning	Platform to train the data and obtain the result of the detection
IV. Edraw Max	Construct basic diagram.

V. Notepad ++	Platform to write a code for the system.
VI. Microsoft Word	Software to complete a documentation of the project.
VII. Microsoft Excel	Software to sort and arrange the data into an attribute and instance.

4.6.2 Hardware Requirement.

Personal Computer (Laptop)

Laptop used as workstation to develop this project which is the system implementation and report documentation. Table 4.3 summarize the specifications of the laptop.

Table 4.6: Personal computer Specification

Specification	Description
CPU	1.7 GHz Intel Core-i5-4210U
Operating System	Windows 10
RAM	4GB
Hard Drive Size	1000GB
Hard Drive Speed	5,400rpm
Hard Drive Type	Serial ATA
Display Size	14
Graphics Card	Intel HD Graphics 4400
Video Memory	Shared
Wi-Fi	802.11b/g/n
Bluetooth	Bluetooth 4.0
Touchpad Size	4.1 x 2.5 inches
Ports (excluding USB)	SD card slot Headset HDMI Ethernet VGA USB 3.0 USB 2.0
USB Ports	3

Card Slots	SD memory reader
Warranty/Support	One-year International Travelers Limited Warranty
Size	18.7 x 12.05 x 2.56 inches

4.7 Conclusion

In conclusion, analysis and design phase is very important phase to consider. With this chapter, we have a better understanding the flow process of the project so that it is use the correct step to run this project. Besides, this chapter includes the software and hardware requirement as it is a part that needs to be implementing. The designs in this chapter also will guide the next phase implementation.

CHAPTER V

EXPERIMENT, ANALYSIS AND RESULT

5.1 Introduction

Based on the previous chapter, the flow and design and also the requirements for hardware and software needed of this project has been clearly stated. In this chapter will show the experiment going after through the process of extraction of data. The result of the experiment will obtained and the analysis are come out from that result. The comparison of the result of data also will be analyze to identify the relevant of the data. Then the software development environment setup involved in this project also will be explained in this chapter.

5.2 Software Development Environment setup

For the implementation of this project can be identified in this section. In this project we use a Windows platform as the recommended operating system. The hardware that is utilized in this project is a personal workstation with standard performance as long as it can run the command in the command prompt and the command can produce a result. This project is developed using Eclipse EE Neon to build a source code, Java Development Kit or we called it as

JDK contains the software and tools needed to compile, debug, and run applications. JDK includes the Java Runtime Environment (JRE), an interpreter/loader (java), a compiler (javac), an archiver (jar), a documentation generator (javadoc) and other tools needed in Java development.The Figure 5.1 shows the software development environment setup that involve in this project.

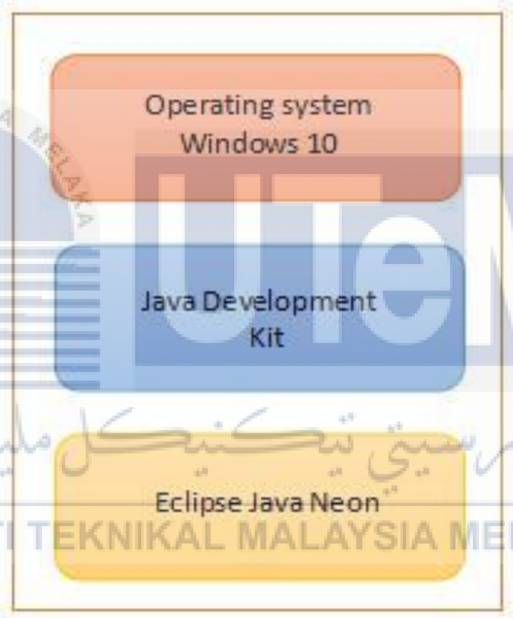


Figure 5.1 : The Software Development Environment Setup Architecture

5.3 Process Module

In this project, there are two main modules involving in this project as shown in figure 5.2 in below.

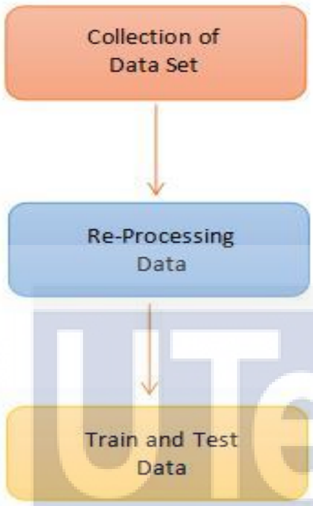


Figure 5.2 : Three Module involve in Frequent Item Mining

Based on the figure above, there are three modules involve in this project, which are collect a dataset, pre-processing data and train and test data. The data has been obtained from the reliable source which contain an image spam. Next module is pre-processing data which mean the data that we get, it must be re-process to know whether the data are include of spam or legitimate. Then, train and test data module will train first consist of same number of instances of image spam and then train it to get the reading accuracy and get the scale model. After that, test data will come out with the predict file. The output of this project will obtain the result of the detection accuracy for Frequent Item Mining which is Maximum Weightage.

5.3.1 Collection of Dataset

For this module, We got this dataset from a previous researcher who had previously created this project which is a image spam. The researcher get this dataset from an internet in website https://www.cs.jhu.edu . The dataset that we get are contained of spam and legitimate in the certain image. This experiment will determine whether the image are consist of spam or legitimate using Frequent Item Mining method.

5.3.2 Pre-Processing Data

Data pre-processing is an important step in the projects. Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. For example, which removes noise from data, normalization, which organizes data for more efficient access.

In this chapter, we have got the dataset already in processed. The dataset it has been classify between the spam and legitimate. The dataset has also determined how many spam and legitimate exist in each file. So, we do not need to do this process again.

5.3.3 Train and Test Data

In this module, the figure 5.3 illustrate the data we get will be train first and then we do the test of the data to get the result. The instances of the image had been chosen of 1000 image to make a random number and it divided into two group which is 500 image for spam and 500 image for legitimate. The total number of instance image are 2300 image. Image number for train and test must be same amount because it will obtain the relevant data.

From all the instances number of image, it will run ten times to get the accuracy result and to making a prediction file by using a feature vector which is Bag Of Word, Minimum Weightage and Maximum Weightage and. After we get the result, we can compare of this feature vector which are the higher result.

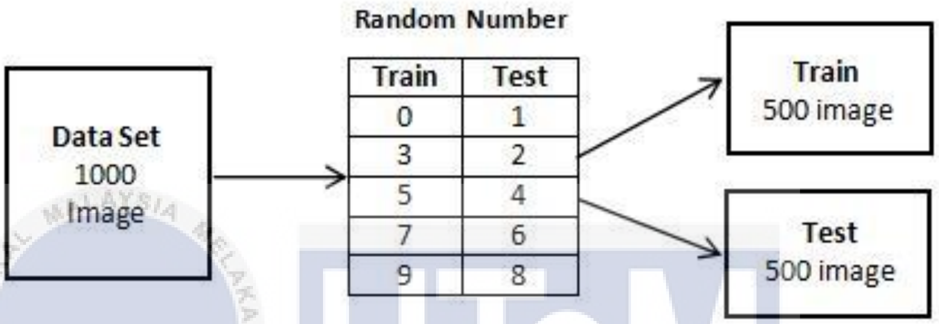


Figure 5.3: Process Train and Test

5.3.3.1 Process Create The train .scale.model

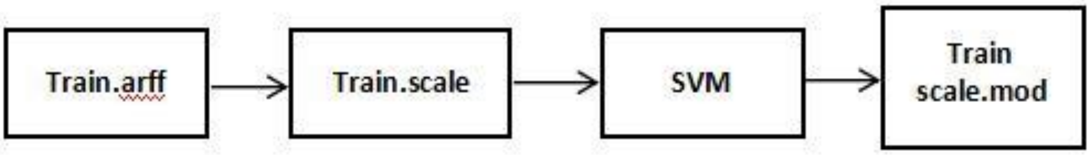


Figure 5.4: Process Create The train .scale.model

Figure above 5.4 shows, from the train .arff file it will used the train .scale and went through the support vector machine(SVM) process to produce the train .scale.model. In train .scale.model have the weightage value that will be used by the test .arff file to generate prediction file.

5.3.3.2 Process Generate the Predict File

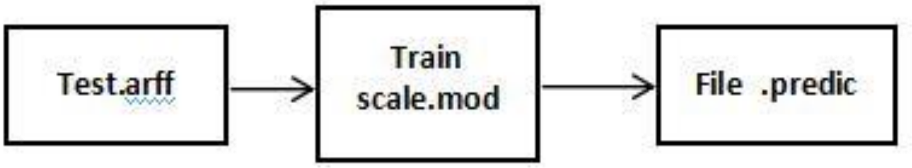


Figure 5.5: Process Generate the Predict File

Figure 5.5 illustrate the test .arff file will used the weightage value in train.scale.model to generate the prediction file. In the prediction file it will notify us wheather it is spam or legitimate using the binary vectors a 0(Legitimate) or 1(Spam). This process from the prediction file it will gives us result with the accurancy of the the BOW(Bag Of Word) and other feature vector such as Minimum Weightage and Maximum Weightage.

5.4 Implementation of Project Execution

To obtain the result of the detection, the program must be compile and run using a jjava code. The command to run the code are shown as below.

```
java -Xmx12G -cp D:\anis\test\predict\Generate.jar azman.PredictFIMAnis
2300 D:\anis\test\predict\ D:\Anis\liblinear-2.11\windows\ "C:\Program Files
(x86)\gnuplot\bin\\"
C:\Users\Administrator\AppData\Local\Programs\Python\Python36\ 0.40 1 10
> resultPM
```

-Xmx4G = RAM used

-cp = execute

Generate.jar = Executable JAR file

azman.PredictFIMAnis = package and source code file name

2300 = Total instances

500 = Number of train instances

0 = Initial number of train instances.

4 = support number

1 = Initial range number of train and test data.

10 = Final range number of train and test data.

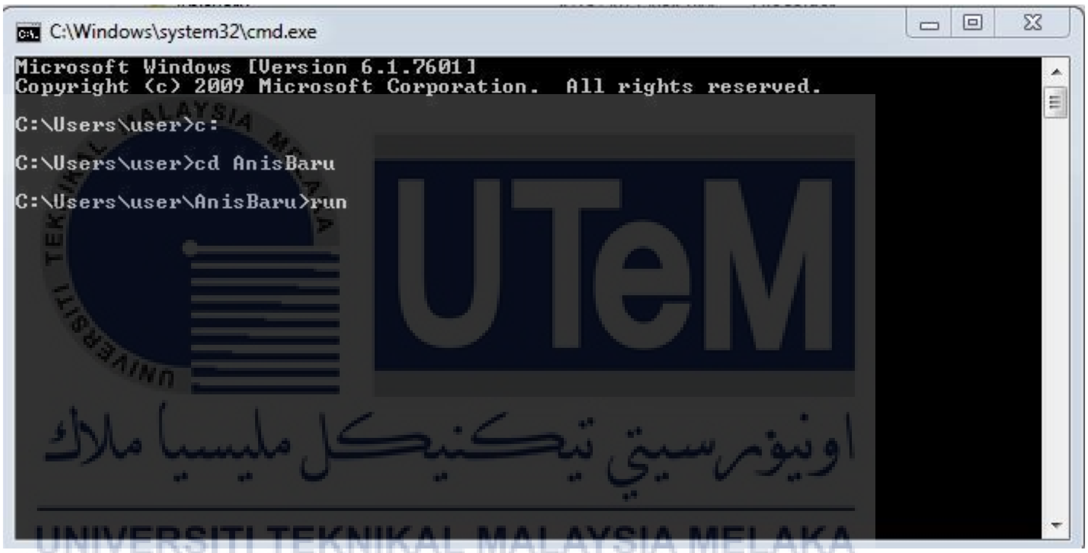
D:\anis\test\predict = Path of source code file.

D:\anis\test\predict\ D:\Anis\liblinear-2.11\windows\ = Path of liblinear file

resultPM = File create for display the result.

This command must save in **.bat** file to run the program easily. Open the command prompt in specified path and just type name of .bat file to run the program.

After execute the process, there are several types of files that will be created in folder such as a arff file, model file, scale file and the predict file for each 10 times run of train and test data in each feature vector. The function of scale model file to make have a data transformation between 0-1 and must follow the formula to make a prediction detection reading. The predict file is a prediction for a test result but it will may have some inaccuracy prediction for the data that have spam and legitimate.

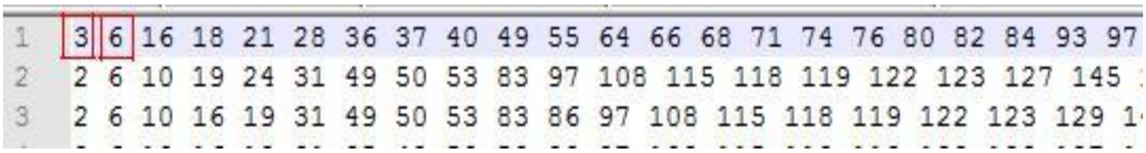


```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

C:\Users\user>c:

C:\Users\user>cd AnisBaru

C:\Users\user\AnisBaru>run
```

Figure 5.6: Run program .bat

To execute the .bat file first open the command prompt and open the local disk of the sourcode and cd for the directory for the folder. To run execute the command used the command run (name of the folder) and click enter to execute the program. After executed , there are several types of files that will be create such as predict file, model file, scale file for each 10 times run of train and test the data in each feature selection.

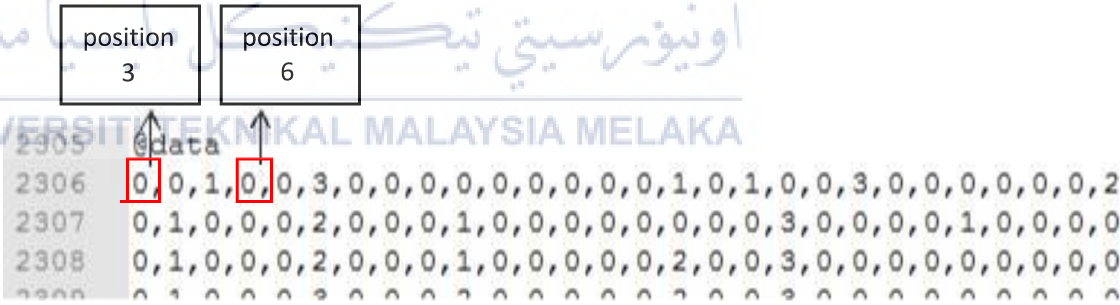
After that, the program will generate a certain rules file such as allRuleSpam and allRuleLegitimate. This file will show the arrangement of spam in file trainBOW. This rule will create a Frequent Item Mining. The explanation will be show at below in Figure 5.7, Figure 5.8,Figure 5.9 and Figure 10.



```
1   3 6 16 18 21 28 36 37 40 49 55 64 66 68 71 74 76 80 82 84 93 97
2   2 6 10 19 24 31 49 50 53 83 97 108 115 118 119 122 123 127 145
3   2 6 10 16 19 31 49 50 53 83 86 97 108 115 118 119 122 123 129 1
```

Figure 5.7: Example of allRuleSpam

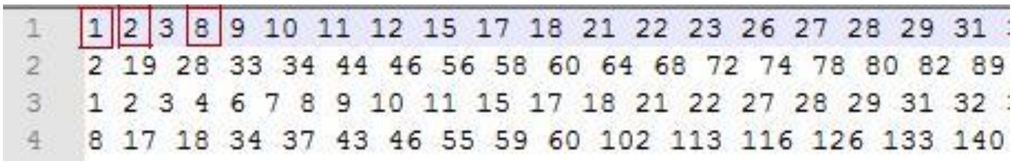
Based on figure 5.7, it show the arrangement of image that include of spam. This is a file allRulesSpam1.file in the folder. Every number in this file are represent a number in file trainBOW-L0-2300-1.arff also in folder.



```
2305  @data
2306  0,0,1,0,0,3,0,0,0,0,0,0,0,0,0,1,0,1,0,0,3,0,0,0,0,0,0,2
2307  0,1,0,0,0,2,0,0,0,1,0,0,0,0,0,0,0,0,3,0,0,0,0,1,0,0,0,0
2308  0,1,0,0,0,2,0,0,0,1,0,0,0,0,0,2,0,0,3,0,0,0,0,0,0,0,0,0
```

Figure 5.8: Example of file trainBOW

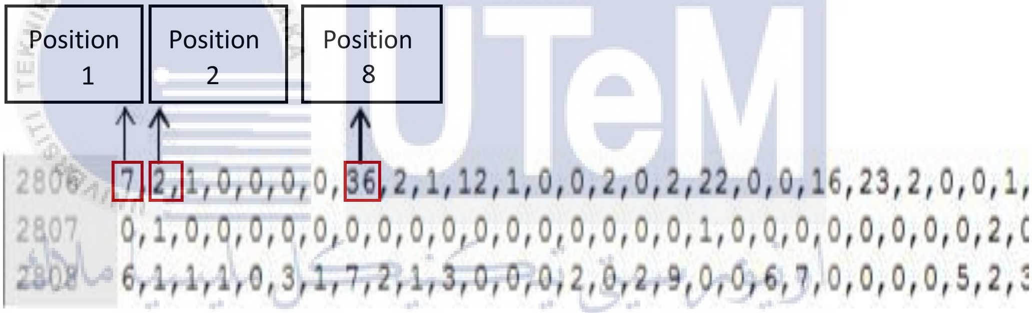
Based on figure 5.8, it illustrate the arrangement of image that include of spam. This file shows the number at allRuleSpam will represent at this trainBow file. For example, look at the number 3 and 6 at allRuleSpam file, then look at the position of the file trainBow. There was a same position. In file trainBOW, the spam exist from line 2306 until line 2805.



```
1  1 2 3 8 9 10 11 12 15 17 18 21 22 23 26 27 28 29 31
2  2 19 28 33 34 44 46 56 58 60 64 68 72 74 78 80 82 89
3  1 2 3 4 6 7 8 9 10 11 15 17 18 21 22 27 28 29 31 32
4  8 17 18 34 37 43 46 55 59 60 102 113 116 126 133 140
```

Figure 5.9: Example of allRuleLegitimate

Based on figure 5.9, it show the arrangement of image that include of legitimate. This is a file allRulesLegitimate1.file in the folder. Every number in this file are represent a number in file trainBOW-L0-2300-1.arff also in folder.



```
2806  7,2,1,0,0,0,0,36,2,1,12,1,0,0,2,0,2,22,0,0,16,23,2,0,0,1,
2807  0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,2,
2808  6,1,1,1,0,3,1,7,2,1,3,0,0,0,2,0,2,9,0,0,6,7,0,0,0,0,5,2,
```

Figure 5.10: Example of file trainBOW

Based on figure 5.8, it illustrate the arrangement of image that include of spam. This file shows the number at allRuleSpam will represent at this trainBow file. For example, look at the number 3 and 6 at allRuleSpam file, then look at the position of the file trainBow. There was a same position. In file trainBOW, the spam exist from line 2306 until line 2805.

5.5 Installation of required software

For this project, all the source code was build and run fully through the Eclipse Java Neon. It has been installed in personal computer with Windows 10 operating system. This software will succeed all the source code steps involved in this project which is a feature vector and to train and test the dataset whether it spam or legitimate. The software must be installed with Java Development Kit in order to let the operating system run this program.

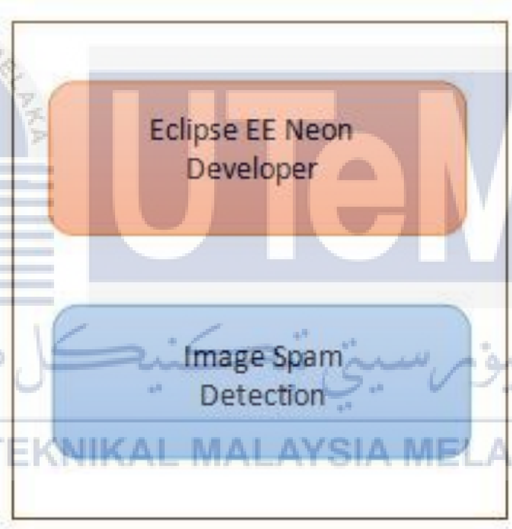


Figure 5.11: Deployment Diagram

The Figure 5.6 shows the arrangement of software and hardware to run the Auto Virus Removal version 1 (AVRv1). The personal computer is a user and will be conducted as required hardware.

5.6 Experiment Result

After run the program and through all the file involved, table below show the result and average of the detection of accuracy. The table is arranged by the Bag Of Word, Minimum Weightage and Maximum Weightage. For the Minimum and the Maximum Weightage has been divide into three level which is for Minimum Weightage have Pick One 1, Pick One 2, Pick One 3 and Naive Pick One. For the Maximum Weightage have a Pick Many 1, Pick Many 2, Pick Many 3 and Naive Pick Many. For the result BOW and NPO are the existing results that have been done by the previous researcher. The table 5.1 shows the result of the ten times of program run.

Table 5.1: Result and average for BOW,Minimum and Maximum

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
1st	87.70%	85.90%	85.90%	83.90%	83.80%	73.30%	73.00%	85.60%	86.50%
2nd	90.80%	88.50%	88.50%	87.70%	82.40%	75.30%	75.60%	90.10%	89.50%
3th	90.10%	85.90%	85.90%	85.30%	85.80%	73.00%	73.00%	86.00%	86.40%
4th	88.70%	85.70%	85.70%	81.80%	80.20%	50.00%	50.00%	85.80%	86.10%
5th	91.30%	86.60%	86.60%	86.50%	84.80%	65.70%	65.70%	86.40%	86.60%
6th	87.00%	82.70%	82.70%	81.60%	80.90%	50.00%	50.00%	83.20%	81.50%
7th	88.20%	85.00%	85.00%	83.70%	82.70%	70.70%	70.60%	85.60%	86.10%
8th	90.00%	86.00%	86.00%	86.10%	84.50%	80.20%	80.80%	87.20%	85.90%
9th	91.30%	84.90%	84.90%	83.30%	82.90%	69.50%	69.50%	85.00%	85.10%
10th	90.00%	86.20%	86.20%	83.30%	83.30%	70.40%	70.40%	86.80%	86.10%
Average	**89.51%**	**85.74%**	**85.74%**	**84.32%**	**83.13%**	**67.81%**	**67.81%**	**86.17%**	**85.98%**

5.6.1 Bar Graph for First Run

Table 5.2: Result for First Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
1st	87.70%	85.90%	85.90%	83.90%	83.80%	73.30%	73.00%	85.60%	86.50%

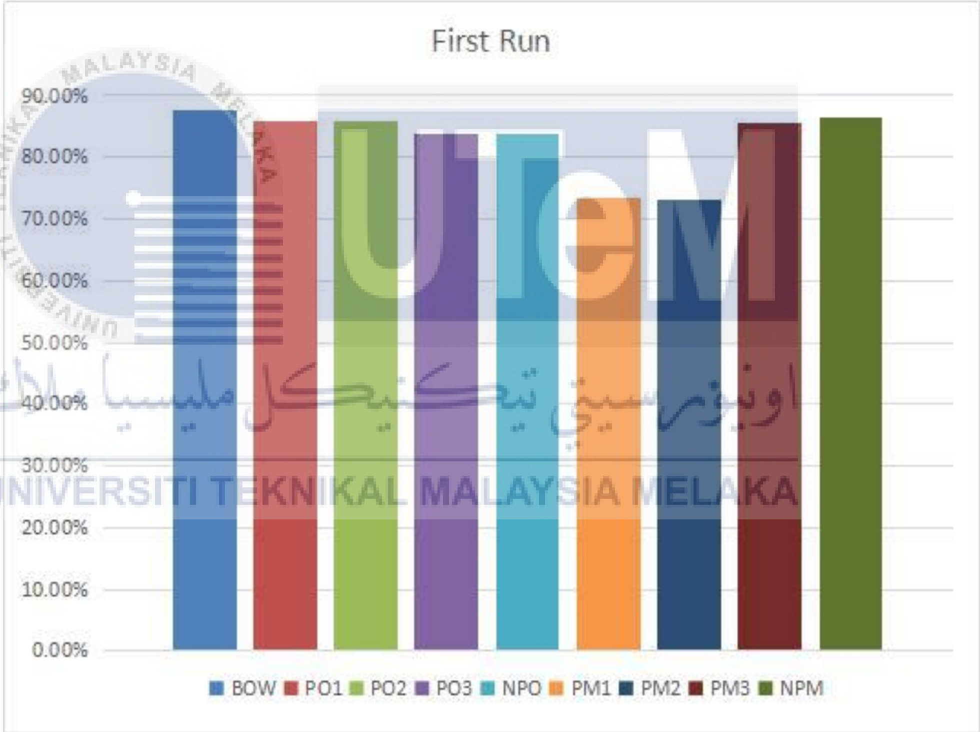


Figure 5.812: Result for First Run

Based on the graph 5.1 at first run, as we can see percentage for BOW is 87.70% is highest from all the other feature vector. The second highest is NPM with 86.50%. Next, the PO2 is represent of 85.90% same as PO1. Then the value of PM3 is 85.60%. For the feature vector PO3 is 83.90%. NPO represent only a 83.80%. Last and but not least, PMI consist of 73.30% and the lowest is PM2 because it represent only 73.00%.

5.6.2 Bar Graph for Second Run

Table 5.3: Result for Second Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
2nd	90.80%	88.50%	88.50%	87.70%	82.40%	75.30%	75.60%	90.10%	89.50%

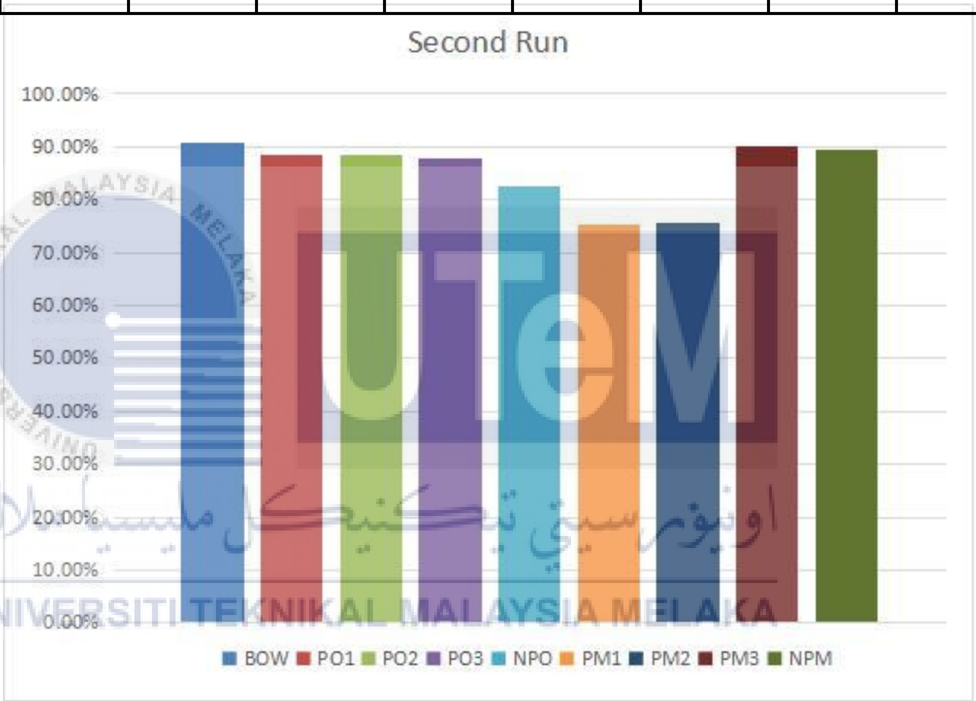


Figure 5.13: Result for Second Run

Based on graph 5.2, the highest result is BOW with 90.80%. Then the second highest result is PM3 which is 90.10%. Next, result for NPM is 89.50%. The result for PO2 and PO1 are same which is 88.50%. Result for the PO3 is 87.70%, result for NPO is 82.40% and for the PM2 is 75.60%. For the last lowest result is PM1 with percentage 75.30%.

5.6.3 Bar Graph for Third Run

Table 5.4: Result for Third Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
3th	90.10%	85.90%	85.90%	85.30%	85.80%	73.00%	73.00%	86.00%	86.40%

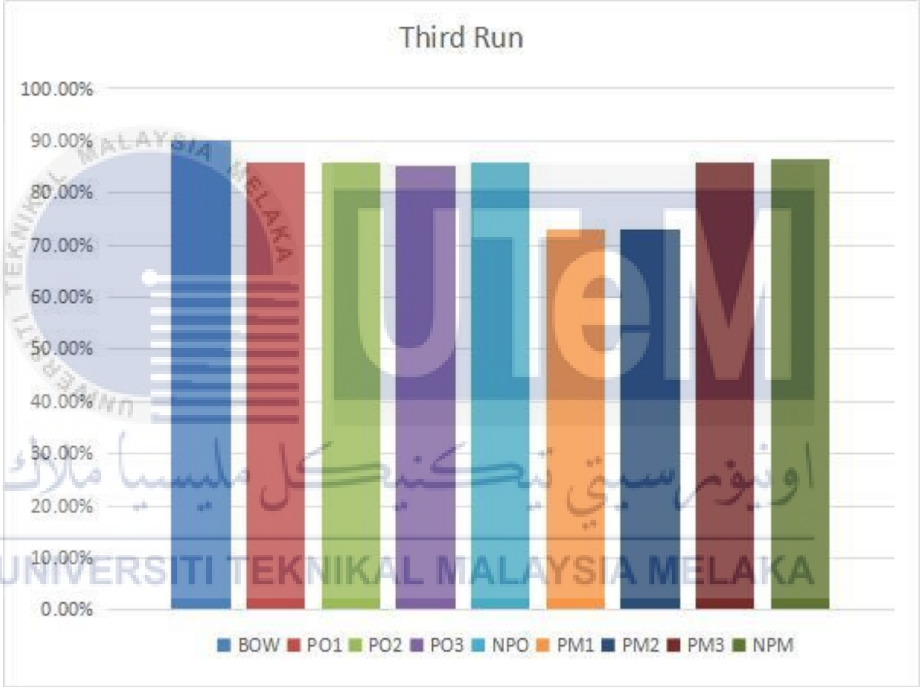


Figure 5.14: Result for Third Run

Based on graph 5.3, the highest percentage is BOW which is 90.10%. Then the NPM is 86.40%. Next, result for PM3 is 86.00%. For PO2 and PO1 is the same value which is 85.90%. The result for PO3 is 85.30%. Lastly, result for PM2 is 73.00% and also for PM1 is 73.00%.

5.6.4 Bar Graph for Fourth Run

Table 5.5: Result for Fourth Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
4th	88.70%	85.70%	85.70%	81.80%	80.20%	50.00%	50.00%	85.80%	86.10%

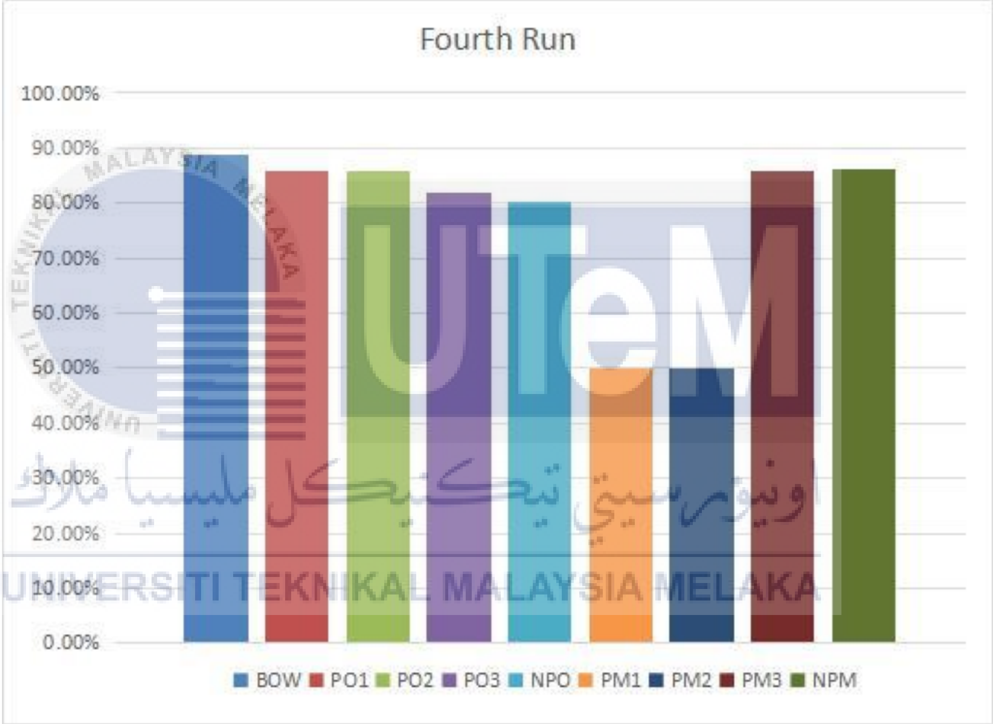


Figure 5.15: Result for Fourth Run

Based on graph 5.4, the highest result is BOW with 88.70%. Then the second highest result is NPM which is 86.10%. Next, result for PM3 is 85.80%. The result for PO2 and PO1 are same which is 85.70%. Result for the PO3 is 81.80%, result for NPO is 80.20% and for the PM2 and PM1 have a same percentage which is 50.00%.

5.6.5 Bar Graph for Fifth Run

Table 5.6: Result for Fifth Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
5th	91.30%	86.60%	86.60%	86.50%	84.80%	65.70%	65.70%	86.40%	86.60%

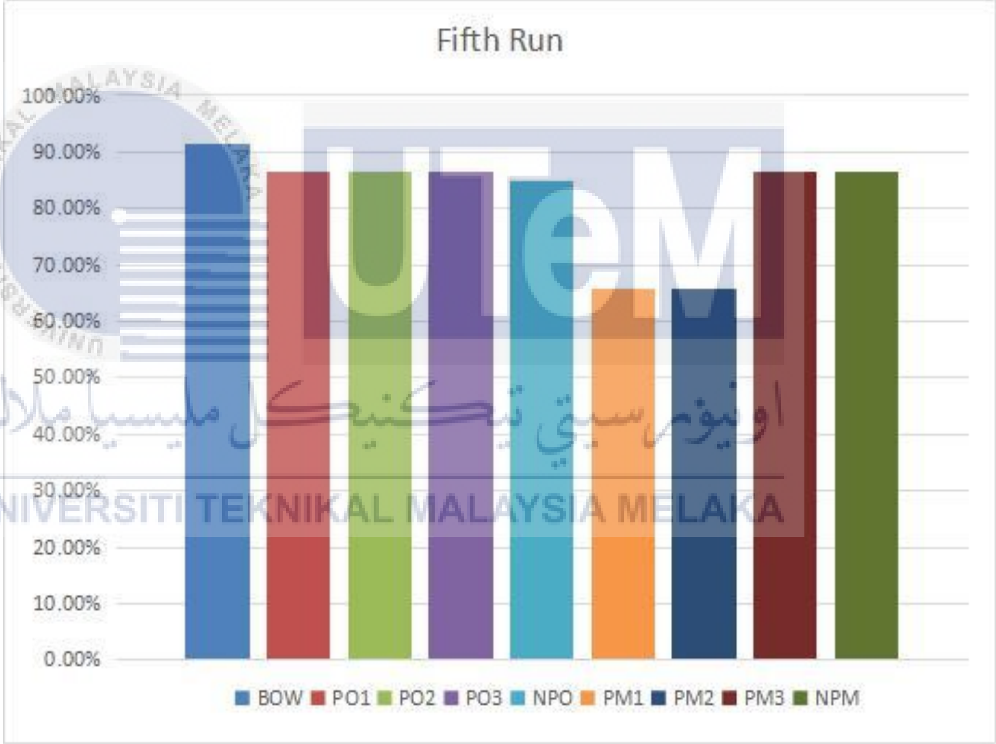


Figure 5.16: Result for Fifth Run

Based on graph 5.5, the highest result is BOW with 91.30%. Then the second highest result is NPM, PO2 and PO1 with percentage 86.60%. Then, for PO3 is 86.50%. Result for PM3 is 86.40% and result for NPO is 84.80%. Lastly, result for PM2 is 65.70% are same with PM1 65.70%.

5.6.6 Bar Graph for Sixth Run

Table 5.7: Result for Sixth Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
6th	87.00%	82.70%	82.70%	81.60%	80.90%	50.00%	50.00%	83.20%	81.50%

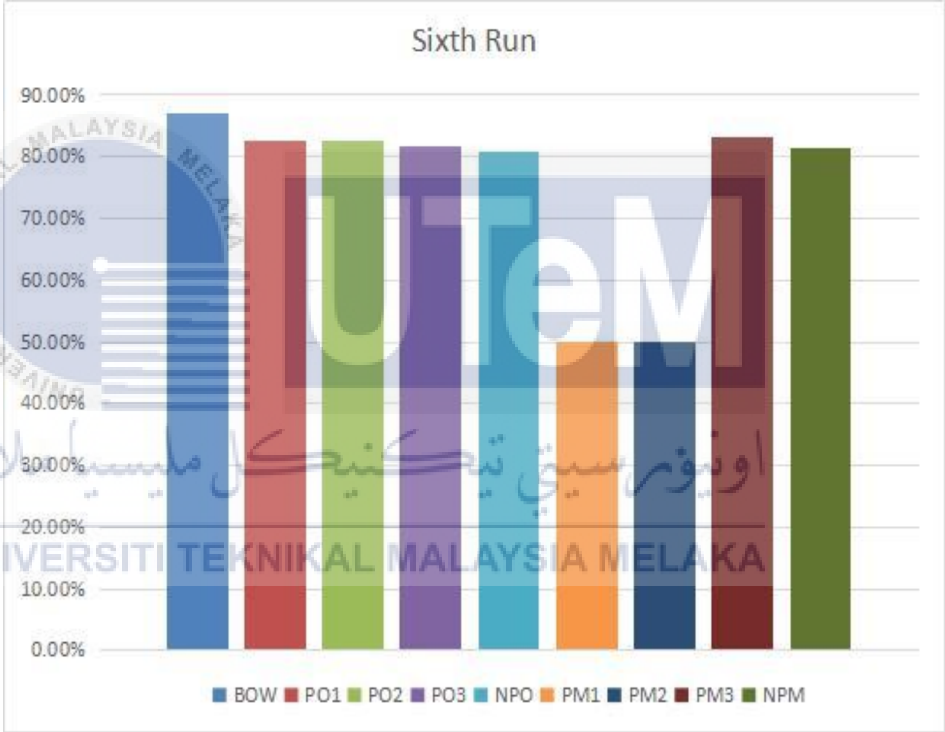


Figure 5.17: Result for Sixth Run

Graph 5.6 shows that the highest percentage is BOW which is 87.00%. For the second highest result is PM3 is 83.20%. Next, result for PO2 and PO1 are same which is 82.70%. Then, result for PO3 is 81.60% and result for NPM is 81.50%. For NPO percentage is 80.90. Lastly, for PM2 and PM 1 is the same percentage which is 50.00% only.

5.6.7 Bar Graph for Seven Run

Table 5.8: Result for Seven Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
7th	88.20%	85.00%	85.00%	83.70%	82.70%	70.70%	70.60%	85.60%	86.10%

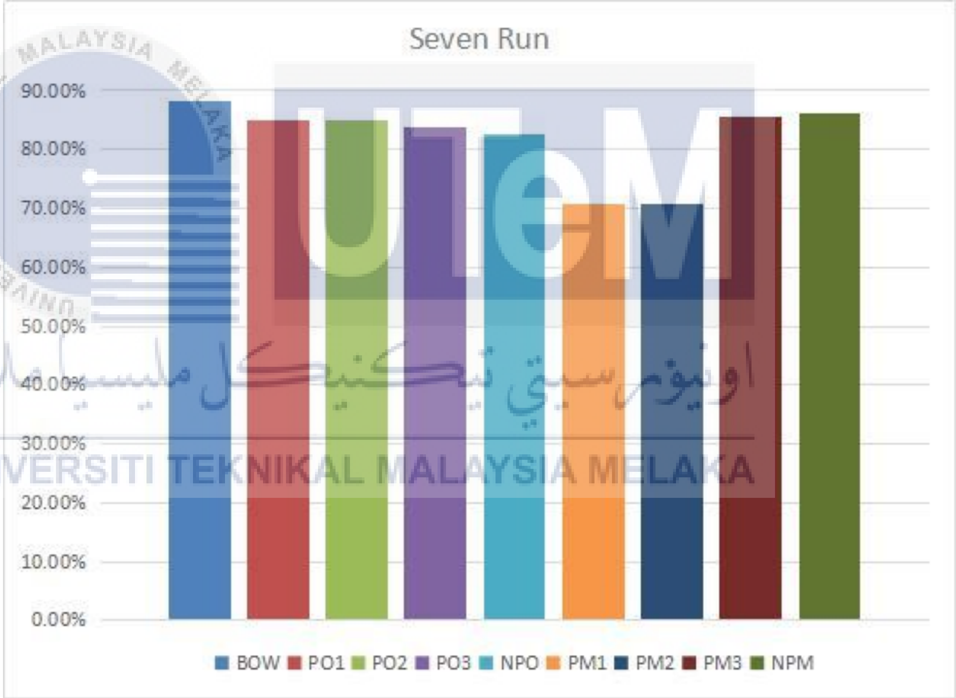


Figure 5.18: Result for Seven Run

Based on the graph 5.7, as we can see the highest percentage on eight run is BOW 88.20%. Result for NPM is 86.10% and for PM3 is 85.60%. For the PO2 and PO1 is the same percentage which is 85.00%. Then, for the PO3 is 83.70%. Next, NPO is represent a 82.70%, PM1 is about 70.70% and the last is PM2 is 70.60%.

5.6.8 Bar Graph for Eight Run

Table 5.9: Result for Eight Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
8th	90.00%	86.00%	86.00%	86.10%	84.50%	80.20%	80.80%	87.20%	85.90%

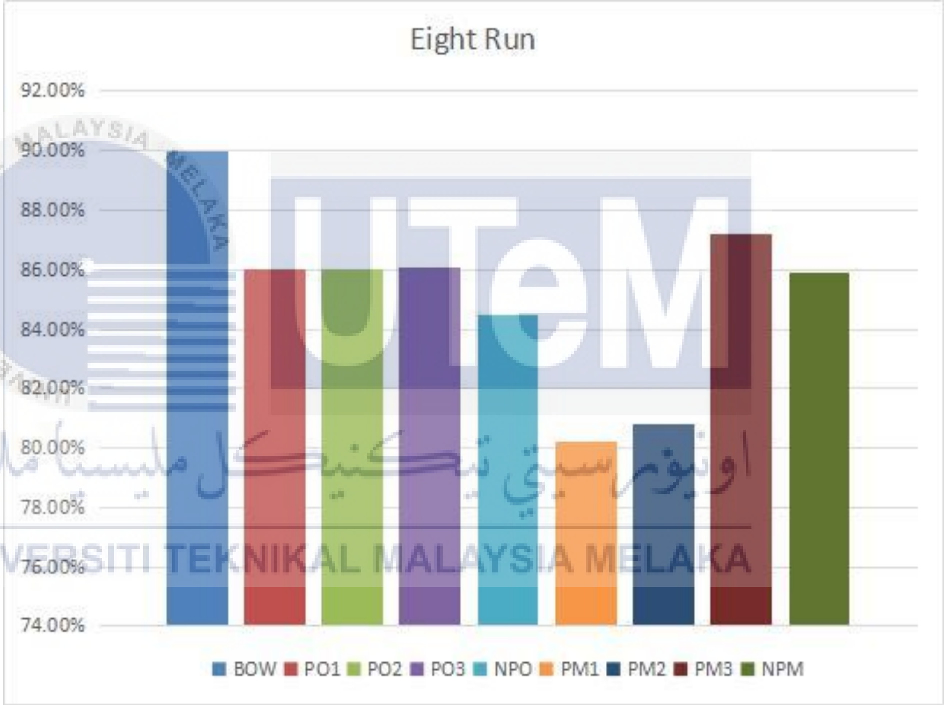


Figure 5.19: Result for Eighth Run

Based on the graph 5.8, as we can see the highest percentage on eight run is BOW 90.00%. For the second highest result is PM3 with percentage 87.20%. Then, result for PO3 is about 86.10%. Result for the PO2 and PO1 is 86.00%. Result for NPM is 85.90%, for NPO is 84.50%,PM 2 is 80.80 and last for PM1 is 80.20%.

5.6.9 Bar Graph for Nine Run

Table 5.10: Result for Nine Run

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
9th	91.30%	84.90%	84.90%	83.30%	82.90%	69.50%	69.50%	85.00%	85.10%

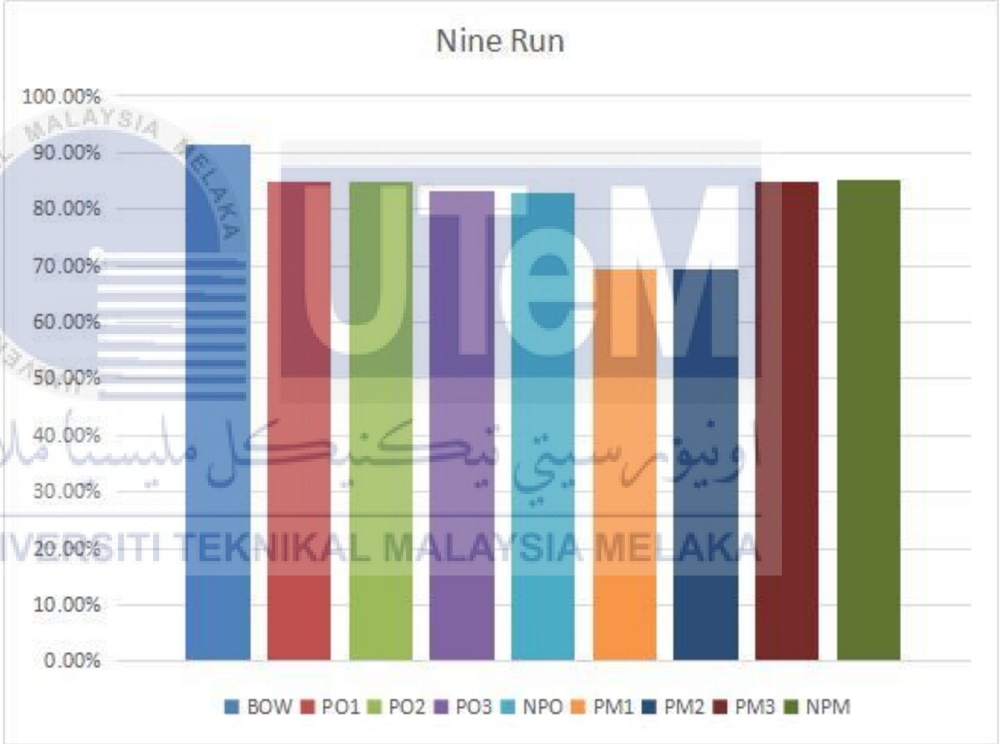


Figure 5.20:Result for Nine Run

Based on graph 6.9, we can conclude that the highest value is BOW which is 91.30%. For the second highest is NPM is 85.10%. Result for PM3 is 85.00%. Next, result for PO2 and PO1 are same which is 84.90%. Then for the PO3 is 83.30% and for the NPO is 82.90%. Lastly, result for PM2 and PM1 is the same value which is 69.50.

5.6.10 Bar Graph for Ten Run

Table 5.11: Result for Ten Run

Num of run	**BOW**	**PO1**	**PO2**	**PO3**	**NPO**	**PM1**	**PM2**	**PM3**	**NPM**
10th	90.00%	86.20%	86.20%	83.30%	83.30%	70.40%	70.40%	86.80%	86.10%

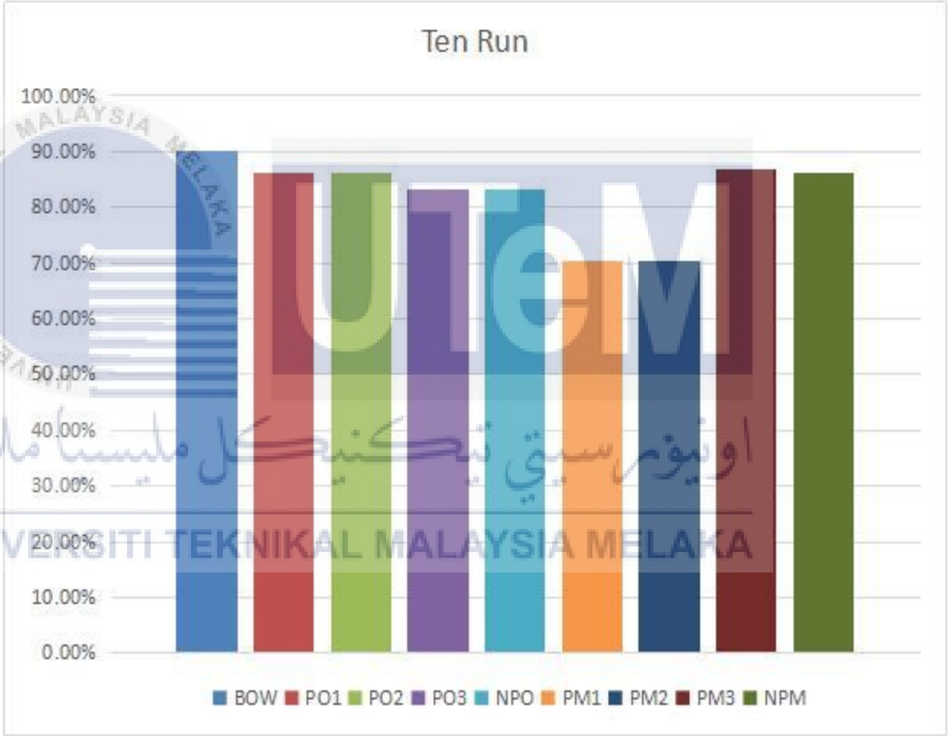


Figure 5.21 Result for Ten Run

Based on the graph 5.10, the greatest percentage is for the feature vector BOW with 90.00%. For the second greatest result is 86.80%. Result for the PO2 and PO1 is same which is 86.20%. The result for the NPM is 86.10%. Next, the result for the NPO is also same as PO3 with percentage 83.30%. Lastly, result for the PM2 andPM1 also same which is 70.40%.

5.6.11 Bar Graph for Average of each Feature Vector

Table 5.12: Result for Average of each Feature Vector

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
Average	**89.51%**	**85.74%**	**85.74%**	**84.32%**	**83.13%**	**67.81%**	**67.81%**	**86.17%**	**85.98%**

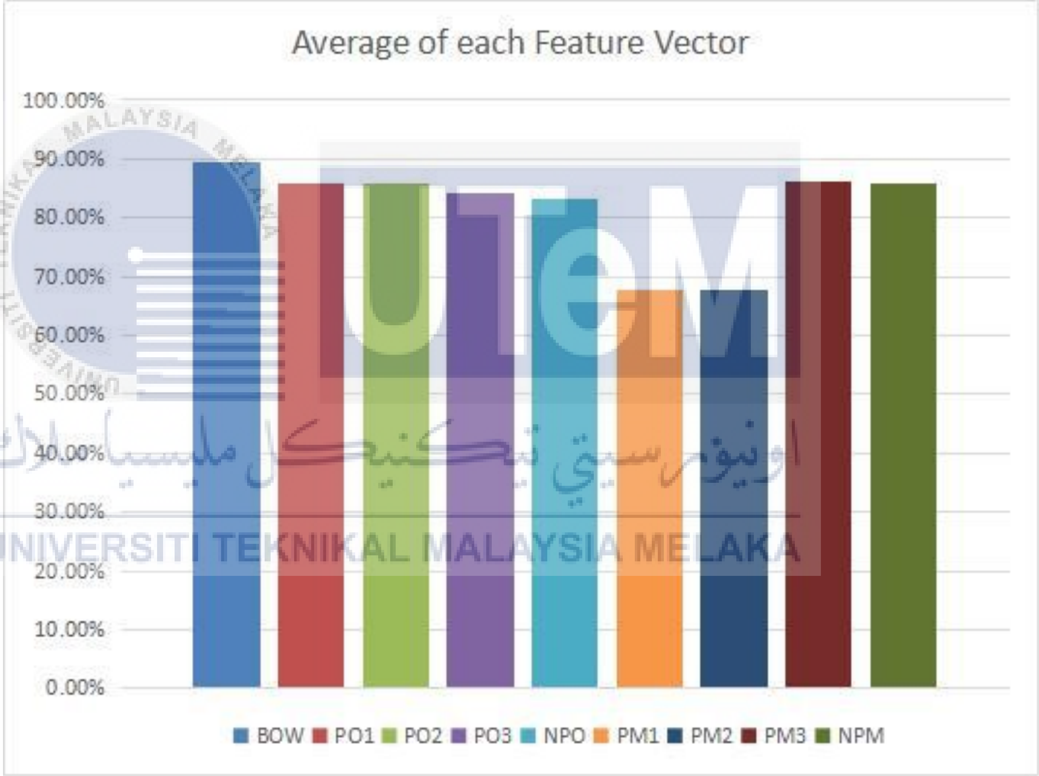


Figure 5.22: Average Result for Each Feature Vector

Refer to the graph 5.11, the highest result is BOW 89.51%. The second highest is PM3 is about 86.17%. Third is result for NPM is 85.98%. For the PO2 and PO1 it have a same value which is a 85.74%. Next, For PO3 is 84.32% and for NPO is 83.13%. Lastly, the lowest percentage is PM2 and PM1 which is a 67.8% only.

5.7 Conclusion

In conclusion, this chapter describes about the process to get the result. In this case, with the dataset, we can generate some of feature vector. Train and test also important to produce an accurate results. From the results that we get, we create a bar graph to explain more about each feature vector for ten times run.

CHAPTER VII

DISCUSSION

6.1 INTRODUCTION

The study of the different image spam detection techniques reveals that image spam detection is a challenging task. The spammers are smarter than the spam detectors. Image spam detection can be seen as a various way comprising of machine learning techniques, classification techniques, artificial intelligent techniques and also image analysis. The selected papers were available freely on line. We reviewed the techniques for image spam detection. The techniques found in the literature reviewed were classified based on the type of features of image spam that were used to detect image spam. We found that the classification accuracy is affected by the feature selection and feature vector. In this chapter we will discuss about the comparison result of new research and previous research. This chapter also Will explain the advantages and disadvantages of our research.

6.2 Comparison of Experiment Result

Table 6.1: Average Result for All Feature Vector

Previous Researcher

Num of run	BOW	PO1	PO2	PO3	NPO	PM1	PM2	PM3	NPM
Average	89.51%	85.74%	85.74%	84.32%	83.13%	67.81%	67.81%	86.17%	85.98%

This project and the previous project use the same classifier which is a Support Vector Machine (SVM). The SVM is to produce a feature vector that has been code in the program. SVM classifier with kernel function is used to identify an image spam and also the accuracy will be calculated.

Based on the table 6.1 above, we can conclude that a Bag Of Word (BOW) made by previous reseacher is higher percentage than all existing feature vector including a Naive Pick One (NPO). In this case, all feature vector can not be beat with BOW even all the feature vector are run in ten times. NPO is a combination from PO1, PO2 and PO3,

For our new research, we have created some of feature vector which is a Pick One 1(PO1) known as Level 1, Pick One 2(PO2) known as Level 2, Pick One 3(PO3) known as Level 3, Pick Many 1(PM1) known as Level 1, Pick Many 2(PM2) known as Level 2, Pick Many 3(PM3) known as Level 3 and Naive Pick Many (NPM). NPM is a combination from PM1,PM2 and PM3.

However, for the final result, all the new feature vector that are created can not beat with the Bag Of Word but Pick Many 3 (PM3) can beat the Naive Pick One (NPO) average result. This is because the different feature vector give a representation to an image that can lead to a better prediction performance.

6.3 Conclusion

For conclusion in this chapter, we have made a comparisons between the results of new research with the previous researcher. In this chapter, we also identify that the same dataset will produced a different result when used a different technique.

CHAPTER VII

CONCLUSION

7.1 Introduction

This chapter explains about the project conclusion. It discusses about overall general information of this project such as contribution and limitation of this project. Future works of this project are also stated so that the project can be enhanced and improved by others. In this chapter, the project conclusion will be contributed. The project summarization, project contribution, limitation and future research will be discussed in this chapter. The project summarization will summarized the whole project. The weakness and proposition for improvement will be discussed in the project limitation and the future research..

7.2 Project Summarization

In view of early part of this project, we have set up to achieve three objectives which are to investigate the performance of the classifier using the feature vector generated from weightage schemes of maximum, to identify the performance which is the best between a single classifier and to identify the performance of ensemble method when combined of these model.

The first objective has been achieved in chapter 3, where we have generate a maximum weightage from BOW feature vector. For the second objective, has been achieved in chapter 6, where we compare the result BOW, Minimum Weightage and Maximum Weightage. While for the third objective, we does not achieved the objective which is to identify the performance of ensemble method when combined of these models. But sometimes in a project we have weakness and advantages respectively.

7.2.1 Observation of Weakness and Strength

Based on observation, there is a weakness and strengths of this project. Besides, the weakness and strengths are important to identify because when we know about the weakness of this technique, we will learn more and make an improvement in future. The table 7.1 will explain more about the advantages and disadvantages of the project.

Table 7.1: Weakness and Strength of Project

WEAKNESS	STRENGHT
The result that we get cannot beat with result of Bag Of Word (BOW)	The result that we get are outperform from Minimum Weightage
Does not achieved one objective	Achieved a two objective

7.3 Project Contribution

This project contribution is useful to public because it performed by a new technique, which is using a Frequent Item Mining. This project are proposed a Maximum Weightage to generate a list of 2-pairs or more of the frequent items. The cluster of 2-pair or above frequent items will selected from a pool of frequent items which consists of the sets {C2-C3}, {C3-C4}, and {C4-C5}. This method maybe will help SVM classifier better.

7.4 Project Limitation

For the project limitation, the objective for identify the performance of ensemble method when combined of model cannot be achieved because we do not have an enough of time to run this experiment. We spend a lot of time while making an experiment for Maximum Weightage.

Other than that, We are not exposed to the collection of dataset process because the dataset we get is a dataset that has been processed. In this case, This is because we have to learn the process from the beginning to the end for get the accuracy result

All the datasets are only run on PCs in CCNA Lab and sometime this program code is successfully run and sometime unsuccessful. This is causes we do not have a time to making an experiment.

7.5 Feature Works

In order to improve the Frequent Item Mining, We will try to use the mean weightage to get the an average for each cluster in one image. When used this way, we will able to see which of result for feature vector is more better.

7.6 Conclusion

In conclusion, the project has achieved two objectives and scopes defined in the Proposal and Chapter 1. Throughout the run of this project need more adaption in order to have better performance and can be implemented and widely used in Frequent Item Mining.

REFERENCES

Abdolrahman Attar , Reza Moradi Rad, Reza Ebrahimi Atani. 2011. A Survey Of Image Spamming And Filtering Techniques

Battista Biggio, Giorgio Fumera, lgnazio Pillai, Fabio Roli. 2011. A Survey And Evaluation Of Image Spam Filtering Techniques

Enrico Blanzieri, Anton Bryl. 2009. A Survey Of Learning-Based Techniques Of Email Spam Filtering

EVANSTON, ILLINOIS. 2010. Machine Learning For Image Spam Detection From Server To Client Solution

d. Saiful Islam[1], Abdullah Al Mahmud[2], and Md. Rafiqul Islam. 2010. Machine Learning Approaches For Modelling Spammer Behaviour

Lamia Mohammed Ketari, Munesh Chandra, Mohammadi Akheela Khanun. 2012. A Study Of Image Spam Filtering Techniques.

Thiago S. Guzella *, Walmir M. Caminhas. 2009. A review of machine learning approaches to Spam filtering

Wanli Ma, Dat Tran, Dharmendra Sharman. Detecting Image based Email Spam