YOUTUBE SPAM CLASSIFICATION USING WORD FREQUENCIES

MEGAT MUAZAM BIN MEGAT THARIH AFENDI



2019

DECLARATION

I hereby declare that this project report entitled

[YOUTUBE SPAM CLASSIFICATION USING WORD FREQEUNCIES]

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

without citations.



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

DEDICATION

All glory and thanks to Allah for paving the way for all that I have achieved.

This thesis was entirely dedicated to my beloved parents who are my source of inspiration and provide great moral, religious, emotional and caring support.

Finally, my family, friends and colleagues, who help me to finish and always give me the opportunity to complete this report.



ACKNOWLEDGEMENTS

Above all, thank Allah S.W.T for the blessings and encouragement that allow this work to be carried out.

I want to thank Mr Nor Azman bin Mat Ariff most warmly for his valuable work in helping me to design and develop this research project. He was greatly appreciated for his willingness to give his time very kindly.

I would also like to thank my supportive parents for their ongoing support, stability and affection, which keeps me inspired and focused all the time.



ABSTRACT

YouTube is among the largest websites and has been one of the Internet's most popular sites. Recognizing YouTube's features is indeed crucial for network activity and to sustainable development of this new service generation. Spam are seen as the most rapidly growth attacks that have infected lots of users all around the world especially in YouTube. In this study will be use YouTube spam collection data set that obtain from UCI Machine Learning Repository website which is this data are from among the 10 most viewed on the collection period and frequently use by past researcher. This dataset process through Bag-of-Word, Chi-Square, and Information Gain to propose SVM model that produce from this project. The objective of this project is to prove that SVM can provide result accuracy in detecting spam and ham comment on YouTube website. The project is giving hope to produce a system that can distinguish between spam and ham comment on web site methods based on SVM model.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

ABSTRAK

YouTube adalah diantara laman web yang terbesar dan merupakan salah satu laman web yang paling popular di internet. Bagi aktiviti rangkaian mengenalpasti ciriciri youtube adalah amat penting untuk pembangunan mampan generasi perkhidmatan baru ini. Spam dilihat sebagai serangan paling cepat yang menjangkiti banyak pengguna di seluruh dunia terutama di YouTube. Dalam kajian ini akan menggunakan set data koleksi spam YouTube yang diperoleh dari laman web UCI Learning Learning Repository yang mana data ini adalah dari antara 10 yang paling banyak dilihat pada tempoh pengumpulan dan sering digunakan oleh penyelidik masa lalu. Proses dataset ini melalui Bag-of-Word, Chi-Square, dan Information Gain untuk mencadang model SVM yang dihasilkan dari projek ini. Objektif projek ini adalah untuk membuktikan bahawa SVM dapat memberikan ketepatan hasil dalam mengesan komen spam dan ham di laman web YouTube. Projek ini memberikan harapan untuk menghasilkan sistem yang boleh membezakan antara komen spam dan sah pada kaedah laman web berdasarkan model SVM.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

TABLE OF CONTENTS

DECLA	ARATION	п
DEDIC	ATION	III
ACKN	OWLEDGEMENTS	IV
ABSTR	ACT	\mathbf{V}
ABSTR	RAK	VI
TABLI	E OF CONTENTS	VII
LIST C	FTABLES	XII
LIST C	OF FIGURES	XIII
LIST C	ويور سيتي تيڪني FABBREVIATIONS	XV
LIST Q	FATTACHMENTS NIKAL MALAYSIA MELAKA	XVI
CHAP	TER 1: INTRODUCTION	1
1.1	Introduction	1
1.2	Problem Statement (PS)	1
1.3	Project Question	7
1.4	Project Objective	8
1.5	Project Scope	8
1.6	Project Contribution	9
1.7	Thesis Organization	9

PAGE

	1.7.1	Chapter I: Introduction	9
	1.7.2	Chapter II: Literature Review	9
	1.7.3	Chapter III: Project Methodology	10
	1.7.4	Chapter IV: Analysis and Design	10
	1.7.5	Chapter V: Implementation	10
	1.7.6	Chapter VI: Discussion	10
	1.7.7	Chapter VII: Project Conclusion	10
1.8	Conclu	ision	10
CHAI	PTER 2:]	LITERATURE REVIEW	11
2.1	Introdu	action	11
2.2	Genera	I Categories of Internet Security Attack (ISA)	12
	2.2.1	ISA Definition	12
	2.2.2	Denial of Service Attack	13
	2.2.3	Phishing Attack	13
	2.2.3.1	اونيونر سيتي تيڪنيڪBhishing	13
	2.2.3.2	Spear Phishing KAL MALAYSIA MELAKA	14
	2.2.3.3	DNS Base Phishing	14
	2.2.4	Spam Attack	14
	2.2.5	Malware	14
	2.2.6	Virus	15
2.3	Classif	ication of ISA	15
	2.3.1	Passive Attack	15
	2.3.2	Active Attack	16
2.4	Spam		16
	2.4.1	Spam Definition	16

viii

2.4.2	Spam Type	17
2.4.2.1	E-mail	17
2.4.2.2	Web Spam	17
2.4.2.3	SMS Spam	17
2.4.2.4	Image Spam	18
2.4.2.5	YouTube Spam	18
2.4.3	Spam Detection Technique	19
2.4.4	Spam Analysis Technique	21
Machine	Learning	22
25.1MAL	Machine Learning Definition	22
2.5.1	Datasat	22
2.5.2	Dataset	23
2.5.3	Data Preprocessing	23
2.5.3.1	Definition	23
2.5.3.2	اونيوم سيتي تيڪني Preprocessing Type	23
U ^{2.5.4} EF	Feature Extraction AL MALAYSIA MELAKA	25
2.5.5	Data splitting and Validation	27
2.5.5.1	Cross validation	27
2.5.5.2	Bootstrapping	28
2.5.5.3	Random Subsampling	29
2.5.5.4	Kennard Stones	30
2.5.5.5	Kohonen Neural Networks	31
2.5.6	Feature Selection	32
2.5.6.1	Feature Selection Definition	32

2.5

	2.5.6.2	Feature Selection Type	33
	2.5.7	Classification	37
	2.5.8	Classification Type	37
	2.5.8.1	Generative	37
	2.5.8.2	Discriminative	38
2.6	Critical	Review	39
	2.6.1	Previous Research on Spam	39
	2.6.2	Previous Research on YouTube Spam	42
2.7	Conclus	sion	48
CHAI	PTER 3: F	PROJECT METHODOLOGY	49
3.1	Introdu	ction	49
3.2	Method	lology	49
	3.2.1	Previous Research	50
	3.2.2	اويتور سيني تيڪ Information Gathering	50
	-3.2.3 UNIVE	Define Scope	51
	3.2.4	Design and Implementation	51
	3.2.5	Testing and Evaluation of Model	51
	3.2.6	Documentation	51
3.3	Project	Schedule and Milestones	52
	3.3.1	Project Flowchart	52
	3.3.2	Project Milestones	53
	3.3.3	Project Gantt chart	54
3.4	Conclus	sion	54
CHAI	PTER 4: A	ANALYSIS AND DESIGN	56
4.1	Introdu	ction	56

X

4.2	Problem Analysis		56
4.3	Project	Design	57
	4.3.1	Dataset	58
	4.3.2	Data Preprocessing	59
	4.3.3	Feature Extraction	63
	4.3.4	Generate BOW feature vector	63
	4.3.5	Data Splitting and Validation	64
	4.3.6	Feature Selection	65
	4.3.7	Normalization	66
	4.3.8	Classification	67
4.4	Require	ement Analysis	74
			74
	_4.4.1 ⊨	Software Requirement	/4
	4.4.2	Hardware Requirement	75
4.5	Conclu	sion	76
REFE	RENCES	اونيومرسيتي تيكنيكل مليسيا	77
APPENDIX/ERSITI TEKNIKAL MALAYSIA MELAKA 87			

xi

LIST OF TABLES

Table 1.1 : 2019 YouTube comment removal	6
Table 1.2 : Problem Statement	7
Table 1.3: List of Project Question	7
Table 1.4: Relation of each PS, PQ and PO	8
Table 2.1: SPAM literature	41
Table 2.2: YouTube SPAM literature	47
Table 3.1: Milestone	54
Table 4.1: Description of Dataset.	58
Table 4.2: Advantage and disadvantage of rules	63
Table 4.3: Confusion Matrix of Hyperplane A and B	70
Table 4.4: SVM Kernel	72
Table 4.5: Software Requirement AL MALAYSIA MELAKA	75
Table 4.6: Hardware Requirement	76

LIST OF FIGURES

PAGE

Figure 1.1: Reported Spam Incident 2019	2
Figure 1.2: Reported Spam Incident 2018	3
Figure 1.3: Reported Spam Incident 2017.	4
Figure 1.4 : Web attack sources by country, November 2017 – October 2018.	5
Figure 2.1: Literature Review's Structure	12
Figure 2.2: Example comment spam on YouTube	19
Figure 2.3: Pipeline of different approaches	26
Figure 2.4: Divided sub sets	28
Figure 2.5: Architecture of RSE	30
Figure 2.6: Filtering Method	34
Figure 2.7: Wrapper Methods	35
Figure 2.8: Embedded Scheme IKAL MALAYSIA MELAKA	35
Figure 2.9: Generic Scheme	36
Figure 2.10: Classifying spam comments of recent trends	44
Figure 2.11: Results based two techniques used	45
Figure 2.12: First data set	46
Figure 2.13: Validation data set	46
Figure 3.1: System Framework	50
Figure 3.2: Project Flowchart	52
Figure 4.1: Project Design	57
Figure 4.2: Example of Kate Perry dataset	58
Figure 4.3: Example of features will be use	59
Figure 4.4: Raw data	59
Figure 4.5: Tokenization	60

Figure 4.6: Token Length	60
Figure 4.7: Case Normalization	60
Figure 4.8: Special Character	61
Figure 4.9: Stop word removal	61
Figure 4.10: Stemming	62
Figure 4.11: Feature Descriptor	63
Figure 4.12: Feature Vector	63
Figure 4.13: Subsampling	64
Figure 4.14: 2-Class SVM Graph	67
Figure 4.15: Hyperplane Placement	68
Figure 4.16: Hyperplane with Margin	68
Figure 4.17: Choosing the Best Hyperplane	69
Figure 4.18: Graph with an Outlier	70
Figure 4.19: Non-Linear SVM's Case	71
Figure 4.20: SVM with Multi-Dimensional Space	71

اونيۈم سيتي تيڪنيڪل مليسيا ملاك UNIVERSITI TEKNIKAL MALAYSIA MELAKA

LIST OF ABBREVIATIONS

FYP	-	Final Year Project
ISA	-	Internet Security Attack
DOS	-	Denial of Service Attack
DNS	-	Domain Name System
Email WALAYSIA	-	Electronic Mail
Web	22	World Wide Web
SMS	- 5	Short Message Service
GIF	-	Graphic Interchange Format
ASCII	-	American Standard Code for Information
"Allin		Interchange
XML Jun Jun	~. \S	Extensible Markup Language
SVM-RFE	- 0	Support Vector Machine Recursive Feature
UNIVERSITI	TEKN	Elimination LAYSIA MELAKA
HTML	-	Hypertext Markup Language
IDE	-	Integrated Drive Electronics
Weka	-	Waikato Environment for Knowledge Analysis
BOW	-	Bag-of-Word
SVM	-	Support Vector Machine
FS	-	Feature Selection
IG	-	Information Gain
CS	-	Chi-Square

LIST OF ATTACHMENTS

Appendix A	Gantt Chart Part 1	87
Appendix B	Gantt Chart Part 2	88
Appendix C	Gantt Chart Part 3	89



CHAPTER 1: INTRODUCTION

1.1 Introduction

Social Media is an online community that builds upon Web 2.0 theoretical and computing foundations and facilitates the creation and distribution of user-generated content. (Wolf, Sims, & Yang, 2017). All social media, whether mobile or stationary, involves some kind of digital platform. But not all that is digital, it is typically used to refer to a new media types involving online involvement where email and YouTube has been the most popular form of everyday social networking (Harvey, 2014). YouTube is among the largest websites and has been one of the Internet's most popular sites. Recognizing YouTube's features is indeed crucial for network activity and to sustainable development of this new service generation (Jackman & Roberts, 2014).

1.2 Problem Statement (PS)

A tricky person or group of tricky people or "zombies" is an automated program created to send unsolicited messages in bulk via email or comment. Spammers may be individuals, e-marketers or malicious gangs that are organized into spam networks (Krishnamurthy, 2015). Spammers, however, could also automate spamming by creating computers with specially programmed scripts and viruses. Spammers spam essentially to earn money by inducing various products and services through online marketing. Nonetheless, ham business messages could be tagged as spam in certain circumstances. A genuine product e-brochure could be viewed by the recipient as spam simply because the message came to user without their consent. These are the spammers ' two general reasons where it can disables the email or server of the recipient by sending large messages where spam attackers can paralyze servers or individuals followed by online fraud and phishing via fake pyramid schemes leading to the disclosure of personal information such as credit card number, telephone number, bank account number to defraud the recipient from user cash.

Spam are seen as the most rapidly growth attacks that have infected lots of users all around the world especially in YouTube. Malaysia Computer Emergency Response Team (MyCERT), a well-known established company that provides a point of reference for the Internet community in Malaysia to deal with computer security incidents stated in their Reported Incidents based on General Incident Classification Statistics 2019, by month January to October 209, the number of total spam up to 100 cases. The data derived from MyCERT database which gathered by MyCERT incident report submission. The numbers are expected to steadily growth as the trend in Figure 1.1 shows that it keep increasing over past three months (MyCert, 2019).



Figure 1.1: Reported Spam Incident 2019

Source: MyCERT Incident Statistics,2019

Another alarming statistics were for over past 2 years, spam attack had been fluctuated growth as the trend in Figure 1.2 and Figure 1.3 shows that it at around 344 to 342 reported incident. If we look at the trends over time we can see that the spam attack remained constant with only slightly levelled off by 2 incident.



Reported Incidents based on General Incident Classification Statistics 2018

Figure 1.2: Reported Spam Incident 2018

Source: MyCERT Incident Statistics,2018

In 2018, the number of incident report stood at approximately at 342 report. From January to March there was a slightly increase in incident of 28 report .The number decreased steadily to 13 reports by July then more sharply to 82 in August. At this point the number of incident report remained high for an average of 36 reports per month for the rest of that year. (MyCert, 2018)



Reported Incidents based on General

In 2017, the number of incident report stood at approximately at 344 report. From January to February there was a slightly increase in incident of 38 report. The number decreased to 24 reports by July then increased steadily to 32 in July. At this point the number of incident report remained high for an average of 26 reports per month for the rest of that year (MyCert, 2017).

Kaspersky, a global security company has released a statistics of web-based attacks that contained malware. In 2018, 1 876 999 691 threats from web resources in several countries around the world are blocked by Kaspersky Lab solutions. Figure 1.4 reveals 92.1& records of antivirus blocked threats from online resources in 10 countries (Kaspersky lab, 2018).





The United States (45.65%), the Netherlands (17.5%) and Germany (11.70%), remain in the top ranks. TOP 10 was left by Sweden, Russia and China; Ireland (0.25%), Luxembourg (1.02%) and Singapore (0.98%) took their seats.

Another alarming issues with media social attack is YouTube. A report by (Google, 2019) stated that YouTube is a vibrant community in which every quarter millions of people post billions of comments. For period Oct to Dec 2018 the total comments removed were 166,370,039 where the comment removed detection for automated flagging was 99.4% (165,429,813) and 99.5% (223,374,780) and human flagging was 0.6% (940,226). The number slightly increased for period July to Sept 2018 the total comments removed were 189,788,777 where the comment removed detection for automated flagging was 99.3% (188,550,503) and human flagging was 0.7% (1,238,274). The number then rapidly increased for period January to march 2019 the total comments removed were 278,281,954where the comment removed detection for automated flagging was 99.3% (276,468,960) and human flagging was 0.7%

(1,812,994). The number rose insanely for period April to Jun 2019 the total comments removed were 537,759,344 where the comment removed detection for automated flagging was 99.3% (533,733,638) and human flagging was 0.7% (4,025,806). Table 1.1 shows the reporting period.

Reporting period	Total comments	Comment removed detection
	removed	
Apr 2019 – Jun 2019	537,759,344	Automated flagging: 99.3%
		(533,733,638)
		Human flagging: 0.7% (4,025,806)
Jan 2019 – Mar 2019	278,281,954	Automated flagging: 99.3%
and the second s	<u>s</u>	(276,468,960)
EK.	KA	
		Human flagging: 0.7% (1,812,994)
660		
Oct 2018 – Dec 2018	189,788,777	Automated flagging: 99.3%
يسيا ملاك	کنیکل ما	ويبوم سنڌي (188,550,503)
UNIVERSIT	TEKNIKAL N	Human flagging: 0.7% (1,238,274)
Jul 2018 – Sep 2018	166,370,039	Automated flagging: 99.4%
		(165,429,813)
		Human flagging: 0.6% (940,226)

Table 1.1 : 2019 YouTube comment removal

Due to this situations for the past 1 year statistics, it is crucial to have to have an effective method in predicting and detecting these comments. YouTube spam spotting is seen as one of the best way to solve the problem. The problem statement for this plan is shown in Table 1.2.

PS	Problem Statement
PS1	As the number of comment removal in the YouTube are rapidly increasing.
	It is crucial to have an effective technique in identifying between ham spam
	comment and spam comment. Another issues is to identify the best machine
	learning technique to be used in processing comment data.

Table 1.2 : Problem Statement

PS1: It is hard to identifying between ham spam comment and spam comment. Spam comment should be identified and detected to separate between the ham comments.

1.3 Project Question

Three project questions (PQ) resulting from the issue statement should be addressed. In this analysis. The project question is summarized as follows:

- I. What is spam comment in social media YouTube?
- II. How spam comment YouTube will be classified?
- **UNIVERSITI TEKNIKAL MALAYSIA MELAKA**

III. Is the machine learning technique give a better accuracy in detecting ham comment and spam comment of YouTube?

Problem Statement	Project Question
	PQ1
PS1	PQ2
	PQ3

Table 1.3: List of Project Question

The focus of this work is on the above. The problem serves as a guide to ensuring that the goal is met throughout the whole investigation process.

1.4 Project Objective

Project Objective (PO) provide the goals for this research. The research objectives are explained as follow:

- I. To study the taxonomy of YouTube spam attacks.
- II. To develop a Machine Learning system that is able to detect YouTube spam comments.
- III. To test and verify the functionality of the tools created to detect YouTube spam comments.

Table 1.3 shows the relation of each PS, PQ and PO in this research.

Problem Statement	Problem question	Project Objective
abl ()		
مليسيا ملاك	يتي بيڪني99	او بيو PO1-
UNIVERSITI TE	PO2 PO3 MALAYSIA	PO2LAKA
101	1 Q2, 1 Q5	102
	PQ1, PQ2, PQ3	PO3

Table 1.4: Relation of each PS, PQ and PO

1.5 Project Scope

The scope of this project (PS) is limited to the following criteria as follow

- I. Dataset obtained from UCI Machine Learning Repository.
- II. The project only focus to Machine Learning method only.

1.6 Project Contribution

The output of this project (PC) is to evaluate the spam comment attack feature relevant to YouTube and select the best feature to boost prediction accuracy. The selected features are used to create effective and reliable tools for spam comment detection together with the machine teaching elements. The overview of this contribution to the project is given below:

- I. Identification of various spam comments attack based on the taxonomy of its attributes.
- II. Propose a model that is able to detect spam comment attacks.
- III. Propose a testing and validation on the developed program to verify the accuracy in detection spam comment attacks.

1.7 Thesis Organization

This report consists of six chapter namely Chapter 1: Introduction, Chapter 2: Literature Review, Chapter 3: Project Methodology, Chapter 4: Analysis and Design, Chapter 5: Implementation, Chapter 6: Testing and Validation, and Chapter 7: Project Conclusion.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

1.7.1 Chapter I: Introduction

This chapter act as the main guide in highlighting the essential items before the project is carried out. It will describe what and how this project tried to achieve and its relevancy.

1.7.2 Chapter II: Literature Review

This chapter encompass on the related studies by previous researcher which later analyzed to find the gap between them thus becoming this project's contributions.

1.7.3 Chapter III: Project Methodology

Chapter III's aim was to construct a rigid set of works that need to be done in order to answer the previously mentioned objectives. The methodology involves the steps and procedures in any possible ways throughout this project.

1.7.4 Chapter IV: Analysis and Design

This chapter analyses and design related procedure that relate to the series of experiments. The experiments are first analyzed its affectivity by referring to the literature and then designed carefully to achieve their respective goals.

1.7.5 Chapter V: Implementation

This core chapter are where all discussed procedures in the previous chapters are been carried out. It will describe in details each steps involved and the environmental setup for the experiments.

1.7.6 Chapter VI: Discussion

This chapter discuss the results and analyze them to show whether the objectives were answered.

1.7.7 Chapter VII: Project Conclusion ALAYSIA MELAKA

This chapter will summarize the project, state the contribution and highlighting the constraint faced throughout the project. This chapter will also describe what next can be done in the future to further improve the project.

1.8 Conclusion

This research are carried out to find out more features of spam comment attacks to improve the process of prediction in distinguishing ham comments and spam comments.. For the prediction fed by the selected features, the machine learning algorithm is used to create a model. The design is then modified and evaluated against other approaches. The following section deals in detail with the related works / literature based on internet security attack and machine learning.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter aim to discuss on the related works regarding internet security attacks and also its classification method as in Figure 2.1. The literature will describe in details on internet security attack categories, classification of internet security attack, machine learning and spam referring from several verified sources. It will act as the critical summary of the related topic of published researched. This will give the clear explanation on what has just been done, what is commonly acknowledged, what is emerging and what is the present state of thinking on the topic. Most importantly, this literature will help in better understanding of research problem being studied.

اونيۆمرسيتى تيكنيكل مليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA



Figure 2.1: Literature Review's Structure

2.2 General Categories of Internet Security Attack (ISA)

2.2.1 ISA Definition

In terms of computer networks, an attack is a trial to steal, disable, kill, change, or obtain unauthorized access or create unauthorized property usage. (S.Mangrulkar, R. Bhagat Patil, & S. Pande, 2014). Although the Internet is the largest computer network in the world, millions of computers are connected. A network can be a cluster of two or many linked computer systems (Gordon, 1995). Protection can also be seen as assured liberty from impoverishment or necessity, safeguards to deter stealing, surveillance, or any problem that ensures or guarantees (Brooks, 2010).

2.2.2 Denial of Service Attack

A Denial of Service Attack (DOS) is a type of network architecture attack that prevents a server from serving its customers. (Elleithy & Blagovic, 2006). An attack by the DOS (Denial of Service) aims to prevent the ham users from approving the access to a network asset or halting systems activity and functionality (Ali, 2006). DOS attacks that are listed as follows, destructive attacks such as removing or altering configuration information or power interrupts, which disrupt the device's ability to act, resource consumption attacks that harm the device's capacity to work, including opening through simultaneous system connections and bandwidth overload attacks that seek to overpower the network device's bandwidth capacity (Fishman et al., 2000).

2.2.3 Phishing Attack

Phishing has originally been a term used to describe email attacks designed to rob your username and a string of characters that allow access to a computer system or service. (Danhieux, 2013). Phishing can be done via text, social media or telephony, but most people are now identifying attacks coming in via email with the word phishing. Phishing emails can penetrate any size and form of organization (goleman, daniel; boyatzis, Richard; Mckee, 2019). Phishing is a source of sensitive data, for instance by taking the appearance of a strong element in electronic mail for example usernames, passwords and charging points of concern cards (Dakpa & Augustine, 2017). There are many type of phishing attack happen in the world such as clone phishing, spear phishing and DNS base phishing.

2.2.3.1 Clone Phishing

This type of phishing attack through which a valid e-mail has its substance in a connection or reference and has served to make an e-mail which is virtually indistinguishable or cloned. The attachment or reference in the e-mail is replaced by a malicious copy and is then forwarded from the original sender. The phisher then cloned the response and sent it to the client (Ma, 2013). A specific variation in the phishing process, which intercepts a ham previously sent email and produces a replica that misleads the target. The victim can easily accept the original email resend and rely on any ties or attachments (Orhan, 2018).

2.2.3.2 Spear Phishing

Spear phishing is an effort to encourage a specially selected victim to open a malicious link or visit a malicious website, in order to gain information on confidential data or to act against the organization of the victim (Larkin, 2005). Spear phishing is the perfect vehicle for a broad range of harmful attacks, from the point of view of cyber criminals. Targeted executives are typically leading officials with names such as CFO, CFO, Senior Vice President and Director of financial affairs. Spear phishing e-mails were produced with sufficient detail to fool even seasoned safety staff (FireEye Inc., 2016).

2.2.3.3 DNS Base Phishing

DNS base phishing attack, when the attacker jeopardizes the domain search process in order to lead the User to a fake website with a click. DNS spoofing is the method of creating a DNS entry that points to another Address (Issac, Chiong, & Jacob, 2006). One of the core protocols for efficient work of web applications is the domain name system. This helps you to link a domain name to its IP address, thus reducing the need for the IP address of a web server to be maintained (Tripathi, Swarnkar, & Hubballi, 2018).

2.2.4 Spam Attack UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Spam is an email you haven't wanted or requested. Spam is the useful newsletter or sales ad for another person. Spam is the usual way viruses, Trojans are propagated. (Gómez, Casas, Inzunza, & Costa, 2017). Spam is unsolicited email that typically advertises rich-fast plans, assignments, loans and porn sites. Spam also comes with fake return data, which makes coping with the perpetrators more difficult (Mob, 1999).

2.2.5 Malware

Malware is short for computer malicious. Technology is used and built to perturb the operation of machines, capture sensitive data and obtain access to private computer systems. Various early infections were written as experiments and buggies, including the first Internet Worm (Gómez et al., 2017). The program, scripts, active material and other programs can appear here. Malware' is a general term for a variety of types of code which are offensive, disruptive and irritable (Ye, Li, Adjeroh, & Iyengar, 2017). Malware today is primarily used for the purpose of stealing sensitive information from others, whether personal, financial or business. Malware is often commonly used to capture and interrupt the activities against government or corporate websites. Malware forms include malware, Trojan, Worms, Spyware, Zombie, Phishing, Spam, Adware and Ransomware (Padmavathi & Divya, 2013).

2.2.6 Virus

A computer virus is a program that, usually without the knowledge of the user, can spread through computers and networks by copying itself (Mob, 1999). A software or code loaded on your computer and that goes against your interests without your knowledge. Also, viruses can replicate. The man is made of all computer viruses. To order to spread to other devices, viruses migrate to other disks. You can just be mischievous or they can make your files extremely destructive, with computer viruses.

(Gómez et al., 2017).

2.3 Classification of ISA

Attack can be marked into two, Active Attack and Passive Attack types. The attack is marked as active when it tries to change or impact system resources, thereby jeopardizing the reliability or quality of the network or network service. A passive attack attempts to locate or generate network information without impacting system resources and thus endangers privacy (S.Mangrulkar et al., 2014).

2.3.1 Passive Attack

A passive attack tracks uncensored traffic and searches for clear text passwords and sensitive information which may be utilized in other attack styles. Passive attack results in the disclosure to an attacker by details or data files are revealed without the consent or knowledge of the user (S.Mangrulkar et al., 2014). Any data that is secret or confidential is the main purpose of this attack. Submitting or obtaining private or public key or any secret information may be this secret information. (Pawar & Anuradha, 2015). It is very difficult to detect passive attacks as such attacks do not impact user traffic or regular network operations. To get some valuable information, the attacker tests the distance, time and frequency of wireless transmission (Khan, Mast, Loo, & Salahuddin, 2008). The main activities of passive attackers are network analysis, weakly encrypted message decoding and authentication data collection. Passive attackers (K. Kumar, 2016).

2.3.2 Active Attack

An active attack includes attempting to circumvent or break down protection, installing malicious code and manipulating or stealing information (S.Mangrulkar et al., 2014). Some of several successful active attacks include spoofing attack, Wormhole attack, Modification, Server Denial, Sinkhole attack, and Sybil assault (Pawar & Anuradha, 2015). In active attacks, disruptive people attempt to access system resources and damage the system and its functions. The types of active attacks may include, Masquerade, Replay, Message alteration, Denial of service attack (DOS) (K. Kumar, 2016). An attacker can repeat old data streams, switch communications or delete selected parts of important communication messages (Shahzad, Pasha, & Ahmad, 2017).

2.4 Spam

2.4.1 Spam Definition UNIVERSITI TEKNIKAL MALAYSIA MELAKA

Spamming is the spread of unwanted, unrelated contents in various fields, such as email, instant messages, internet, and web pages. Spam is the misuse of e-mail systems to indiscriminately send unwanted messages (Hayati et al., 2010). The classical concept of spam is that messages sent to several recipients that have not requested mass messages are not requested. Spam problems are caused by the combination of unwanted and bulk factors, the amount of unwanted messages swamp messengers and drown out the messages recipients want (Michelle, 2019).

2.4.2 Spam Type

2.4.2.1 E-mail

Spam mail is so inexpensive that it is sent to a wide range of users indiscriminately. Spam or non-spam messages should be detected if multiple spam messages are received and their emails will destroy the server. Now, there is a lot of advertising, such as marketing to make money or to sell stuff to transmit falsehoods and misinformation, etc. In fact, HTML mails carry a Web-bug in an address to track who reads the text In addition (Renuka, Hamsapriya, Chakkaravarthi, & Surya, 2011). E-mail Spam costs the ISPs because the bandwidth of service providers passes this burden to consumers when a number of spam e-mails are sent to users of e-mail (Singh, Pandita, Kalyanaraman, & Chhabra, 2018).

2.4.2.2 Web Spam

Web spam can greatly affect the performance of search engine results. Therefore, commercial search engines have a great incentive to identify spam sites efficiently and accurately. Web Spam is categorized into three categories of spam, spam and cloaking links (Castillo, Donato, Gionis, Murdock, & Silvestri, 2007). A spam site is a website used for spamming, or earns a large score from other spam sites. Another concept of spam is any effort to annoy the relevant search engine algorithm or anything not done if search engines were not available (Becchetti, Castillo, Donato, Leonardi, & Baeza-Yates, 2008). Spammers intelligently studied and manipulated the vulnerabilities of these models to produce spam. It can be a tricks for recycling, dumping, spinning, frame stitching. Spammers deceive classification algorithms through the densely connected array of pages (Hans, Ahuja, & K. Muttoo, 2014).

2.4.2.3 SMS Spam

SMS spam is any unwanted or undesirable text message, often for commercial purposes that is sent indiscriminately to your mobile phone. It may take the form of a simple message, a call or text link to a phone, a website link for more information, or a website link for downloading a request (Hidalgo, Bringas, Sánz, & García, 2006). Spam messages are used to market dating services, premium numbers or to sell technology and drugs. It form of spam usually includes schemes and promotions for

different products. This is sometimes also used by service providers to pinpoint the client for some paid activation (Singh et al., 2018).

2.4.2.4 Image Spam

Recently spammers have started to circumvent filters by encoding spam messages as images. This image is usually attached to or inserted in a message whose text includes random words, excerpts from popular literature or even extracts from private non-commercial emails (Yan Gao et al., 2008). Image spam emails are shown in various forms of spam emails. Spam image emails are usually classified into two groups, presents the spammers goal URL to be used in text-driven spammers and a reference to a website (Attar, Rad, & Atani, 2013). Spammers have used photo spammers to bypass text-based spam filters that consist of spam text embedded in an image. Therefore, the issue of spam detection is the distinction between ham and spam images (Annadatha & Stamp, 2018).

2.4.2.5 YouTube Spam

One of the most widely used functionality of Youtube is its comment system which allows users to comment on uploaded videos. This feature enables users to share thought and feelings on the video. Nevertheless, this also allowed malicious users to share spam-related promotional content. Spam commentary is often entirely irrelevant to the video and is created as users in computer bots. Bots are studied in the capacity to execute spam campaigns to post malicious comment on a large-scale (Aiyar & Shetty, 2018). A comment spam example is shown in Figure 2.2 in one of YouTube most watched videos (Alberto, Lochter, & Almeida, 2015).



Figure 2.2: Example comment spam on YouTube

Up to date, YouTube platform has not published any findings on handling malicious users. It only considers text comment as part of spam message. In addition, YouTube announced through its Policy detect spammers, it depends on user's engagement in reporting or flagging at a channel or comment. (Yusof & Sadoon, 2017). An approach may provide a reasonable result, especially when users respond and report on malicious content. Nevertheless, there are also users who abuse it. These users report any dislike video as YouTube spam, hence resulting the topic to be closed immediately, even though their report is not valid. This problem needs to be solved as YouTube is becoming a prominent part of daily life routine (Science, Handayani, Nurmaini, Yani, & Husni, 2017).

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2.4.3 Spam Detection Technique

Spam based on imagery is a spammer-inducing trick that has been implemented a few years ago to embed all text data in a picture and add the picture collected to spam email so that any textual review carried out by a spam filter is avoided. The detection method detects image spam by using low-level object text features, tiny character-specific fragments, the appearance of large character fragments, and large context types overlapping with character (Biggio, Fumera, Pillai, & Roli, 2008). Symantec is considered a leading supplier of spam security products, including spam security products, and offers full database anti-spam and antivirus protection for Symantec Brightmail AntiSpam. Symantec Brightmail built on your website allows unauthorized messages to be deleted without infringing on the privacy of your users before they get into the inboxes (Khawandi, Abdallah, & Ismail, 2019). Official YouTube blog documented efforts in coping with inappropriate feedback by detecting malicious links, identifying ASCII art and making long-term improvements. Another common solution is to automatically block users from spam. But unlike other social networks and e-mails, YouTube spam is typically not created by bots, but uploaded to popular videos by real users who seek to promote themselves spam. Furthermore, the parallels between such messages and legal messages are harder to identify. It also states that cross-validation is not recommended, as previous samples should be used for training the methods and newer ones for testing. Additionally, errors associated with each class during spam filtering should be treated differently as a blocked valid email is worse than an unblocked spam (Alberto et al., 2015). As stated, it is necessary for this research to be done by focusing on text. There are two types of spam detection technique for text spam which are origin based technique and content based spam detection techniques.

For origin based technique it is divided into three parts blacklists, whitelist and realtime blackhole list (RBL). For blacklists achieving fast-path techniques for quickly filtering out reclassified or related spams on incoming social network objects is required to reduce classification costs. Entries are added to the lists because of spam or abuse, and therefore objects containing such entities are likely to be rejected (D. Wang, Irani, & Pu, 2011). Meanwhile for whitelist it include the transmitters (or the domains) for the distribution of incoming mail. Every other mail is deprecated. A whitelist based system would return a sender's request for verification to allow ham senders to meet the recipient and respond within a short time. It is almost certain that no spam can enter the user's inbox if a whitelist is used (Garcia, Hoepman, & Nieuwenhuizen, 2004). Whitelists are opposite blacklists - e-mail lists that are accurate for ham rather than spam delivery (Sosa, 2010). Simply list such approaches. A Whitelist is a list that comprises the user's email addresses or whole domains (Rosi, 2018). The user also uses an automated white list management tool that helps to automatically add known addresses to the whitelist.(Kant, Sengamedu, & Kumar, 2012). Besides for real-time black hole list this method of spam filtering works in the same way as an agreed blacklist for less practical support is required and the Mail Abuse Prevention system (MAS) and system administrators (third-party). In order to authenticate the IP address of the sender against the database, it need to connect to the third-party system whenever an email enters (Singh et al., 2018).
For content based spam detection technique it is divided into three parts rule based filters, Bayesian filter, support vector machine and artificial neural network. For rule based filters based on some filtering rules features to decide if the incoming email is a spam. E-mail, messaging content, keywords, and e-mail header data can be defined by rules (Wu & Deng, 2008). Next is Bayesian filter is an advanced algorithm to filter keywords. In comparison to rule based filtering, Bayesian filtering does not need preset rules and message content analysis. If the likelihood value is greater than the preset threshold, then the messages are marked as spam and treated accordingly, the probability of spam is determined on the basis of the Bayesian principle (Wu & Deng, 2008). After that is support vector machines usually handles template classification, which means that this algorithm is mostly used for the classification of various patterns. Now, numerous patterns are visible such as for Nonlinear and linear (Pradhan, 2017). The main idea behind SVM is to construct an optimal hyper-plane, for linearly separable patterns, that can be used for classification. The main aim is to optimize the scope so that the trends such as the greater the range size, can be identified more correctly (Tripathy, Agrawal, & Rath, 2016). Lastly is Artificial Neural Network is used to identify ham users by permitted users. The intrusion detection system using neural networks is carried out in two phases training so that normal activity and intrusion can be recognized and testing to evaluate the data set is smaller to ensure the network can identify intrusions it is trained to detect than the learning dataset (K.

Kumar, 2016). UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2.4.4 Spam Analysis Technique

The anti-spam group litter blogs offer a range of Adhoc photo spam solutions including blocking GIF images of known senders, blocking all GIF photos, using FuzzyOcr1 and restricting advanced image processing to small images. Many methods try to understand the contents of the objects through embedded text identification, color analysis or the extraction of embedded text. Besides these essential features are File Format ,File Size, Metadata, Image Size, Average Color, Color Saturation, Edge Detection, Prevalent Color Coverage, Random Pixel Tests (Dredze, Gevaryahu, & Elias-Bachrach, 2007). The characteristics to be considered can be grouped into the following category metadata features, color features, shape features ,noise features (Chavda, Potika, Troia, & Stamp, 2018). Four model construction methods may be

listed as wave, animate ,deform and rotate (Z. Wang, Josephson, Lv, Charikar, & Li, 2007).

In recent years numerous researchers have investigated the categorization technique based on machine learning and pattern recognition techniques for the study of the semantic content of electronic mails. In general, text-categorized documents in unstructured ASCII formats or in formats such as HTML are subject to textual categorization techniques. Clearly, it can be used for RFC 2822-based e-mails too (Fumera, Pillai, & Roli, 2006). Some spam companies and legit email companies have to compile email preprocessing with these four algorithm stages email preprocessing software receives email content, it extracts information and then saves the information (feature) extracted into an appropriate database. Next is extract the spam text and ham text from the feature extraction module, and generate function dictionaries and vectors as an input to the selected algorithm. Followed by, spam classification which is to take standard email classification documents for practice, e-mail pretreatment, collect valuable information, save in fixed-format text documents, break the entire document into sentences, extract a spam document vector, and convert it into a fixed-format vector. Lastly, is test performance by doing performance evaluation (Lassification, 2011).

2.5 -Machine Learning UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2.5.1 Machine Learning Definition

Machine learning is how accurate predictions based on past observations can be automatically taught. The way a machine can be trained to carry out certain tasks with the application for machine learning such as pattern recognition, anomaly detection, prediction, neural networks (Nath, Agarwal, & Ghosh, 2016). Two types of machine learning methods are supervised and unsupervised. The data set is marked and thus trained in supervised learning techniques for a reasonable performance that does not need label data and is thus not processed easily in the proper decision-making process for unsupervised (Tripathy et al., 2016). Briefly, in supervised learning needs to be provided with the information input and required performance training. Unsupervised learning training information is not included and the system provides inputs with no desired outcomes just without supervision (Qiu, Wu, Ding, Xu, & Feng, 2016).

2.5.2 Dataset

The databases comprise raw identities or emails that are then indexed in the XML artifacts of the social networks. The XML artifacts are then sent to the social spam detection system, where the three primary intervention elements are protected (D. Wang et al., 2011). Use of YouTube Content API to access the data sets included. The data sets extracted are based on two main features: the channel's popularity and the availability of current comments. Except these two features, there was no other factor. The databases have therefore been selected randomly (not dependent on stars or anything) (Abdullah, Ali, Karabatak, & Sengur, 2018). Several data sets come from the public study repository for UCI and t, to show the efficacy of the method being proposed. To show the effectiveness of the proposed technique, data sets coming from the public repository are selected in a variety of sizes. (Cateni, Colla, & Vannucci, 2014). The considered datasets are the following, the data sets are called Eminem with YouTube id uelHwf8o7_U, total spam 245, total ham 203, and total comment 448 followed by next datasets called Shakira with YouTube id pRpeEdMmmQ0, total spam 174, total ham 196 and total comment 370.

2.5.3 Data Preprocessing KNIKAL MALAYSIA MELAKA

2.5.3.1 Definition

Text Preprocessing a tokenizing procedure is required to obtain all words that are used in a given word, i.e., by eliminating all punch marks and substituting tabs and other non-text characters for a single white space, a text document is broken down into a string of words. For further storage, this tokenized representation is then used. The compilation of different words that are generated by combining all of a collection's text documents is called a data collection dictionary. (ZDH, 2018)

2.5.3.2 Preprocessing Type

As an imaging method, the image enhancement technique is so far better for a particular application than the original image. The raw photographs from the scanning

center and the databases cannot be viewed directly due to several noises found in these objects. Before testing, it should be pre-processed. Conversion transfer, image resize reduction, noise elimination, and quality improvement contribute to an image in which information are correctly found (Perumal & Velmurugan, 2018). Preprocessing consists of a number of transformations from image to image. It does not enhance information about the documents contents, but may help to extract them. The text, language and font classification that is considered to be metadata for recovering content (Kalaskar & Dhore, 2012). Preprocessing is the important method that influences automated detection of defects. Contrast adaptation, frequency modification, histogram equalization, binarization, morphological operation are the methods the preprocessing (S.ANITHA, 2010).

As for text preprocessing, filtering, lemmatization, and stemming, the collection of words describing documents can be reduced by filtering and lemmatizing or stemming procedures, to minimize dictionary sizes and hence the dimension of the document classification within the set. Filtering techniques exclude vocabulary from the dictionary and therefore from the texts. The generic filtering approach is to avoid sorting words. The idea of stopping word processing is to delete terms that provide little or no data on text, such as posts, conjunctures, prepositions, etc. (Allahyari et al., 2017). Words which occur extremely often can be said to have little information to distinguish between documents, and words which are very seldom likely to have no statistical relevance and can be removed from the dictionary. A technique of (index) sentence collection is also required to further reduce the number of words in the dictionary (Sathiaseelan, 2017). Lemmatization methods tend to link verb types to the infinite tense and nouns to the singular structure. In order to do this, though, you have to learn the word type, i.e. delegate that word's position in the text document. Since this method of labeling usually takes quite some time and is still prone to errors, stemming techniques are often used in training. Stemming methods attempt to construct basic forms of words, i.e. to remove the plurals, verbs or other appliances from the nouns. A stem has the same (or very similar) normal set of words. Every term is defined by its stem after the stemming cycle. Established a series of production rules to convert iteratively (English) words (ZDH, 2018).

2.5.4 Feature Extraction

Feature extraction specifies the appropriate shape data of a sequence so that a structured process enables the task of classifying the pattern. The common extraction techniques are Model matching, Deformable model, Transformation of Unitary Image, Fourier definitions, Gradient and Gabor characteristics, Zernike Moments, Unitary Image, Descriptors, Projection Histograms, Contour Chart, Zoning and Geometric Moments invariants (G. Kumar & Bhatia, 2014). Feature extraction is a special form of reduction of dimensionality in the identification of patterns and image processing. The data input is translated into a series of functions called extraction of the element. The process to extract the most important information from raw data is the extraction of features. The extraction function determines the parameter set that correctly and uniquely describes the form of a character (Tian, 2013). Extraction of features: Features such as structure, shape, color, and more are used for defining object contents. Image features can be categorized as content-based retrieval (CBIR) for high-tech components such as graphics, signal and object processing, pattern recognition, computer communication with humans and science of human perception (Heavner et al., 2001).

Lexical features indicate the kinds of words, characters and attributes that the author wants to use, for example number of upper cases or averages length. Syntactic characteristics try to represent the revisers ' writing style and include characteristics such as the number of punctuations or feature words such as "a," "the" and "of."(Crawford, Khoshgoftaar, Prusa, Richter, & Al Najada, 2015). The task of Lexical Simplification (LS) is to perform Text Simplification (TS) in the sense of Natural Language Processing (NLP) by concentrating on lexical information. In order to make it simple, it can be formally described as the task of replacing words in a given sentence without any modifications to its syntactic structure. The phrase "complex words" in the survey generally refers to individual words, although most definitions and methods can also be extended to multi-word phrases and some of the work discussed covers these terms. The term "complex expressions" has been used for these cases. LS is about identifying complex words (hereafter "target words") and finding the best candidate replacement for the target words. Although keeping the sentence grammatical and maintaining its meaning as much as possible, the correct alternative must be simpler.

This is a very challenging task, particularly when different target audiences will have different needs, for example, speakers of different languages will be more or less familiar with different second language vocabulary sub-sets. While LS approaches vary in different ways, most use a very similar sequence of steps to simplify sentences. Description of the function as the steps pipeline shown in figure 2.2. (Silpa & Irshad, 2018)



Figure 2.3: Pipeline of different approaches

An n-gram character is a continuous n word string. The number of n grams that can be produced for a particular document (usually n is 1, 2, 3 or 4) is basically the result of a stream of n characters passing around the text. One character at a time is shifted by the screen. Instead, each n-gram counts the number of occurrences (Wei, Miao, Chauchat, Zhao, & Li, 2009). Bag-of-words is the most widely used template for the representation of text. The approach is to record word events in the message. This produces a space-based vector that fits the terms in the message. N-gram is the most direct method that assigns probability to the terms or string. Regardless of whether the next word is going to be predicted or the entire sentence, the n-gram template is one of the main methods of speech and language management (Ogada, Mwangi, & Wilson, 2015). Character n-grams are capable of capturing nuanced, lexical, syntactic and or structural stylistic details. To date, fixed-length n-gram characters have been used for the classification of authors (Houvardas & Stamatatos, 2006).

2.5.5 Data splitting and Validation

A traditional multivariate process requires a number of separate data sets. For simulation, estimating the parameters of the formula or creation of the neural network, the measurement or training data set is important. A second set of data is often needed to decide when the practice will end or to evaluate how many templates and parameters should be included. The second set of data is generally referred to as a computer control array. A third set of data called a test set to choose the most fitting model is required if several models are produced. In order to evaluate the performance of the final model, a validation data set is necessary. Different data have been shown to be necessary for all these sets of data because otherwise the models and estimates are partial. There are numerous subsampling techniques, which include cross validation, bootstrapping, random subsampling, Kennard stones and kohenen neural network.

2.5.5.1 Cross validation

The data is divided into n equal parts for a n-fold cross validation. The first part is used as a set of test data, and the remainder as a set of calibration data. The second part of the test data is then used, while the remaining part is used for a new calibration. This is repeated n times and the n test data predictions are averaged. It is necessary not to move any awareness of the models from fold to fold. There are no clear rules as to how many cross-validation folds should be used, whereby one specimen is extracted at a time as a simple and clear way of carrying out cross validation. This particular form of cross-validation is referred to as absolute cross-validation, departure or cuts, providing a specific, but reproducible outcome. Nevertheless, the increase in cross validation groups has been shown to lead to lower root average predictive square errors that offer overly optimistic estimates of productiveness.

According to (Iii, 2009) Subset size the primary method of λ tuning, such as K-Fold, to estimate a tuning parameter. Divide the information into K roughly the same 5 sections as in figure 2.4. K – 1 sections of a kth component are estimated by β –k(α), and its error is determined for each k = 1, 2,.



Figure 2.4: Divided sub sets

Formula

$$E_k(\lambda) = \sum_{i \in kth \ part} (y_i - \mathbf{x}_i \hat{\beta}^{-k}(\lambda))^2.$$

Cross-validation error formula

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} E_k(\lambda)$$

It usually uses K = 5 or 10 and makes it smallest and makes it smallest for a range of T values and selects the value of $CV(\lambda)$.

2.5.5.2 Bootstrapping TEKNIKAL MALAYSIA MELAKA

The early design of bootstrap resampling allowed researchers to assess if their findings may have improved if a different random test was used in place and how different could be the outcomes when a prototype was applied to new data. The prevalence of bootstrapping in the area of small data sets has also increased. Bootstrapping is based on sampling to create a calibration set with substitution. The 0.632 bootstrap is chosen in n cases from the configuration array for the most common version, where the same specimen can be selected repeatedly. The samples that have not been selected are then used for the test set.

According to (Fook Chong & Choo, 2011) the average bootstrap variance of the mean is the standard deviation of the median bootstrap. he generic median bootstrap defect, which is simply the normal bootstrap medians variance is where ^M is the median of bootstrap sample i and [^]M is the grand mean of all the bootstrap medians.

Formula

$$SE_{BS} = \sqrt{\frac{1}{1000 - 1} \sum_{i=1}^{1000} (\hat{M} - \overline{\hat{M}})^2} = 3.17,$$

For asymptotic contradictions, bootstrapping is not compromised and could be the only way to estimate failure for extremely small data sets, which can randomly replicate the entire procedure.

2.5.5.3 Random Subsampling

Random sub-sampling known also as multiple holdout or redundant analysis tests are focused on arbitrarily separating the information into subsets whereby the client determines the size of the subsets. Data can be replicated randomly through arbitrary partitions. Random subsampling was shown asymptotically consistent, leading to more pessimistic predictions of the test data in comparison with cross validation, in contrast to a full cross validation procedure. The test data predictions provide a realistic assessment of the experimental verification software predictions.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

According to (Pathical & Serpen, 2010) the purpose behind the Random Subsample Ensemble (RSE) is to break into several minor sub-problems a complicated high dimension issue and thus solve the software complexities of the original problem. Through choosing random function subsets or subsamples from the original set, a high dimension field space is mapped onto a number of lower dimensions. A random subsampling can be used to construct the lower-dimensional projections. The three approaches used in order to generate the random subsamples of the initial operational region random sampling without replacement, random sampling with replacement and random mutually exclusive partitioning.

Random sampling without substitution occurs for subsamples in which a chosen element in a subsample is special. Nevertheless, the same function may appear

in more than one sub-sample with various sub-samples. Random substitution samples allow a single trait in a sub-sample and across the subsample array to be replicated. Random partitioning is an exclusive sub-sample, i.e. the chosen attribute is special in one and all subsamples. The same amount of options can be included in each interface subsample for all three strategies, even if this is another variable factor.



A simulation study was performed on UCI database high-dimensional data sets to determine the optimum sampling methodology for the non-replacement, substitution and reciprocal partitioning sampling method. In the simulation analysis, the results show that the ensemble performs better without the process of sampling than the other two strategies.

2.5.5.4 Kennard Stones

Kennard Stones (KS) algorithm dividing data sets into two subsets. The algorithm ends with 2 experiments, which on the basis of the input variables become the most distinct from each other's. Such 2 samples were extracted from the initial set of data and inserted into the data set for validation. The process is replicated until a test array has exceeded the required number of samples. The advantages of this algorithm are that the measurement samples completely map the measured area of the

variable space and all samples fall within the measured area. Nevertheless, this method can only be used for one test trial, as partitioning of information is special in making the algorithm unusable for a resampling procedure

According to (Saptoro, Tadé, & Vuthaluru, 2012) a suitable subset is chosen by the Classic KS algorithm from a collection of N samples. The algorithm follows a step-by-step process for ensuring a uniform distribution of such a subset throughout the x data space, where new selections are made in areas of the space far from the samples already selected. For this purpose, the algorithm employs Eucledian distanceED p q? x ?, between the x- vectors of each pair ??qp, of samples as shown by the equation below



N is the variable number inx, and M is the sample number and j is respectively the variable in p and q. The collection continues with the taking of a couple of specimens with the greatest distance. The algorithm selects a sample with the minimum distance from any already selected sample at each subsequent iteration. This process is replicated until the required number of specimens is collected.

2.5.5.5 Kohonen Neural Networks

The application of Kohonen neural networks is an interesting approach to dividing a dataset into two subsets. The two-layer networks are unsupervised networks that can be used as a two-dimensional mapping tool. A Kohonen network is built with the full data set for repartitioning. Then the first data set will be chosen for each neuron for the particular number of samples that activated this neuron.

According to (Skuratov, Kuzmin, Nelin, & Sedankin, 2019) the Kohonen layer comprises a number of n linear elements parallel to each other which obtain at their outputs the same number of inputs and the same input x = [x1, x2, ..., xi] matrix. At the j-th linear component, it obtain the signal:

$$y_j = \omega_{j0} + \sum_{i=1}^m w_{ji} x_i,$$

Where ω_{ji} is the synaptic weight coefficient of i-th input of j-th neuron, i is the number of an input, j is the number of a neuron, ω_{j0} is the threshold coefficient. However, it is difficult to use the Kohonen networks for several subsampling runs because it is rather subjective to create different selection rules for samples that excite neurons for an arbitrary number of runs. The user needs to enter data from the data set to the data set.

For summarization, the calculation of the prediction potential becomes overly optimistic if the same data set is used for calibration and verification. Each set of data should also be as large as possible. The greater the calibration data, the higher the model and the larger the validation data, the better the predictively estimation. If many information is available, broad and autonomous representative samples for learning, reporting, screening and verification can effectively be used by partitioning the large sample collection. Typically, data sets of limited size are only available in analytical chemistry since measurements are costly and work intensive.

2.5.6 -Feature Selection ** UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2.5.6.1 Feature Selection Definition

The goal of the function selection is to pick a small subset from the original features by eliminating non-relevant, obsolete and disruptive characteristics as a dimensional reduction software (S. Wang, Tang, & Liu, 2016). Feature selection can be refer to an aspect of the data, as a 'feature ' or ' attribute ' or ' variable. Features are usually specified or selected before collecting data. Characteristics may be discreet, continuous or nominal. (Ladha & Deepa, 2011). The aim of feature selection as a technique for increasing dimension is to pick from the initial a small subset of applicable features by deleting obsolete, unnecessary or noisy features. The selection of features generally results in better learning performance such as increased accuracy, lower computer costs and improved interpretability of models (S. Wang et al., 2016). In the choice of features a subset of functions is selected based on consistency and

significance from the initial set of features. According to its relevance and redundancy, there are four types classified as subsets of features: noisy & irrelevant, redundant & slightly relevant, weak, and non-redundant (Venkatesh & Anuradha, 2019).

2.5.6.2 Feature Selection Type

The controlled selection of features is used otherwise uncontrolled feature selection is appropriate when class labels of the data are available. Class labels are unknown for many applications in data mining which show the importance of uncontrolled selection of features (Jothi & H, 2012). This dissertation discusses the technologies used to choose images to be evaluated using machine learning. The choice was made on the basis of good quality, excellent and special content compared to the other photographs chosen. Two binary classifications, such as content and quality, are taken into account (Lorentzon & Wallenberg, 2017). In the machine learning literature, a great number of function selection or weighting methods were proposed. The key difference is between filter methods that determine a category for the features without concern for the inducer classificatory and wrapping methods which scan the optimal subset for the unique inducer in the subsets of features (Setia & Burkhardt, 2006). The reliefF algorithm is a standard example, and in the case of problems involving dependencies in the function space it can effectively give value of each element. The SVM (SVM-RFE) approach is a standard selector form wrapper where the SVM (Support Vector) system is used as a classifying tool. This approach was used first of all to assess the weighing of the qualified SVM classifier in the gene selection where the rating criteria for different characteristics is calculated. Although the SVM-RFE process always works more accurately in classification, it is usually much more computational (Zhou & Wang, 2015).

Chi Square is a statistical test which measures the occurrence of characteristics in relation to the expected number of occurrences. For Chi-square the independent variables are the attributes and the associated factors are the classes (Subramaniam, Jalab, & Taqa, 2010). Chi-Squared is the standard statistical method, which tests the variations between the distributions predicted if the function event is meant to be independent from the category quality. It is regarded as a statistical test that very low predicted quantities are erratically complied with, which is commonly found in text identification because they have unusual term features and sometimes because they have few good examples of a definition (Lee, Lushington, & Visvanathan, 2011).

The information gain is assigned to the evaluator used to select the information, if the search method is selected by default. The benefit of data is multivalued and the pick attribute metric information is provided to select the attribute that gains the maximum amount of information (Dinakaran S & Dr. P. Ranjit Jeba Thangaiah, 2013). The benefit of data is multivalued and the pick attribute metric information is provided to select the attribute that gains the maximum amount of information. Since information benefit considers each function independent of the other, multiple redundant features that can discriminate between categories cannot be identified. In contrast, it provides a ranking of the characteristics according to the information they get, thus making it easy to select certain features. The baseline choice approach for features was the data gain in this analysis. The performance of any proposed method should be at least equal to the quality of information profit (Houvardas & Stamatatos, 2006).

Filter Methods selects features that are relatively independent from classification based on discriminatory criteria. Some approaches use simple Fisher criteria correlation coefficients. Others take reciprocal information (T-test, F-test) or statistical testing. In earlier filter methods, characteristics were evaluated in isolation and correlations were not considered. Figure 2.6 show the filter methods.



Figure 2.6: Filtering Method

Wrapper methods use the classification system as a black box for the prediction of the subsets of characteristics. Machine-learning groups also extensively discussed wrapper approaches focused on SVM. A reversing system is used to remedy recurrently, the SVM-RFE (Support Vector Machine Recursive Feature Elimination) wrapper method, which is used to cancer research. At each recursive level, the features are rated according to the sum of the goal function reduction. The lower graded function of the test is then removed. Some models also use the same extraction scheme and linear kernel backwards. Figure 2.7 show the wrapper methods.



Figure 2.7: Wrapper Methods

The inducer has its own FSA in the Embedded Scheme (whether real or implied). An example of this integration are the techniques that cause logical conjunctions. This system involves other traditional machine learning methods such as decision-making structures or artificial neural networks (Ladha & Deepa, 2011).



Figure 2.8: Embedded Scheme

Filter models are based on the general characteristics of the training data in order for the selection process to be carried out as a pre-processing step without induction algorithm, with the independence of all predictors and filter methodsFor example, category differences or predictive dependency are chosen using general characteristics of training data. The model is quicker than the wrapper approach and generalizes more, as it operates independent of the induction algorithm. However, it tends to select subsets with a large number of functionalities including all functions and thus requires a threshold to select a subset. Ranker's Selection Tool Ranker approaches are evaluating attributes with the rating (true or false), number to be chosen and a threshold value to be set to disregard attributes by their individual assessments, in addition with attribute evaluators (RelativeF, Benefit Ratio, Entropry etc.). Several features are omitted owing to the default meaning. To through the set of attributes, use either this method or this count. A filter method in which it is a preprocessing step regardless of the predictor's choice is the classifications variable ranking. The ranking method usually ranks which attributes should be high or low in the given datasets according to the chosen attribute. Ranker gives the evaluator an orderly assessment of the attributes (Dinakaran S & Dr. P. Ranjit Jeba Thangaiah, 2013).

Symmetrically unsure, some substitute classifier measure must be available for the selection mechanism based on filters that weighs the value of a selected feature subset. To order to evaluate the quality of built solutions, the proposed method has used a data mathematical calculation named symmetric instability to through the comparative figures (Dinakaran S & Dr. P. Ranjit Jeba Thangaiah, 2013).

A research by (Cateni et al., 2014) has shown that the hybrid filter-wrapper model takes into consideration a multivariate data collection that contains N observations and K attributes. Figure 2.9 illustrates a generic approach scheme. The original dataset is used with several function selection methods which have a certain significance for each variable.



Figure 2.9: Generic Scheme

2.5.7 Classification

The classification function can be treated as a supervised method where each specific example belongs to a category showing the importance of a particular intent attribute or just the class attribute. Categorical values may be taken up by the objective attribute, each corresponding to a class. The classification task divides the set of examples taken into two separate and comprehensive groups, namely the training set and the test set. The classification process is divided into two phases: training if a classification model is built from the training set and evaluation when the model is tested in the test set. During the training phase, both the indicator and the parameter values in the training set have been used for all cases (S. V. K. Kumar & Kiruthika, 2015).

2.5.8 Classification Type

2.5.8.1 Generative

Generative models that can create practical data samples seem to be an important tool to raise data volume to address such a void. Generative systems are mainly designed to address the underlying distribution of information through the learning of a system which fits training data. Generational models may produce measurable information values with the trained knowledge distribution. Through training a generative model of a small amount of labeled data, which can be used to train the classifier a large amount of labeled data. (Li, Pan, Wang, Yang, & Cambria, 2018)

Generative approach learning each language is to evaluate the language of the expression in the template class-conditional pdfs. Previous probability generative as the samples can produce artificial data points-such as Gaussian, Naïve Bay, Multicultural mixtures, Gaussian, Hidden Markov (HMM), Sigmoid faith networks, Bayesians Gaussian, Hidden Markov (HMM). (Tang, 2012)

Naive Bayes classification has recently gained great popularity and has shown remarkably good performance. Such probabilistic methods How (words in documents) the information is made and propose an assumption-based probabilistic method. Then use a number of examples of training to evaluate model parameters. Use Bayes law, new example is listed and the category most probable to have been placed. Perhaps the most easy and most commonly used classification is the Naive Bayes classification. This models the distribution of documentation by probabilistic processes in every category, given specific situations are separately distributed. Although in many real world applications this so-called "naive bays" assumption is clearly wrong, native bays are strikingly good. (Allahyari et al., 2017)

Hidden Markov Models are a typical techniques of probabilistic that usually do not take into account the expected labels of nearby terms. Hidden Markov (HMM) method are probabilistic models that take this into account. The Hidden Markov model assumes a Markov system in which a label or statement is created based on one or a few previous labels or comments (Allahyari et al., 2017).

A Bayesian network (BN) is a simple acyclic graph and probability distribution for every node in that graph, in view of immediate predecessors.. A network of the Bayes Classifier is a Bayesian network that gives a certain probability distribution across a number of different categorical attributes. It consists of two parts, the acyclic map G, which contains nodes and arcs and tables of probability. The nodes represent attributes while the arcs are based directly. Models can be Sparse BNs (for example, naive models Bayes, and hidden models of Markov), whereas BNs with high complexity can capture them. Therefore, the BNs offer a robust simulation system. (S. V. K. Kumar & Kiruthika, 2015)

2.5.8.2 Discriminative

A discriminative methodology identifies cultural disparities by understanding any vocabulary. Any effort to model the probability processes underlying computational assets is directed towards improved performance of a given task. Logistic regression, traditional neural networks, Nearest neighbor, conditional random fields (CRF) are popular models (Tang, 2012).

The nearest neighbor classifier is a local classifier that utilizes distance-based identification tests. The main idea is that records belonging to the same group are more likely to be similar and comparable to one another on the grounds of the measurements of resemblance. The identification of the sample paper is calculated by the class marks in the training set for related papers. In the training collection the strategy is called knearest identification when we find the nearest neighbor (Allahyari et al., 2017).

Support Vector Machines (SVM) are classification algorithms supervised for learning which are widely used for problem text classification. SVM is a linear classification method. Linear classifiers in text documents are structures based on the value of the linear variations of the properties of the texts. (Allahyari et al., 2017). Vector support machines (SVMs) are effectively tackled in different classification issues and forecast challenges, as originally proposed by Vladimir Vapnik in the field of mathematical learning theory and structurally threat minimization. SVMs were used for many problem pattern recognition and regression estimation, and were used for problems with dependency estimation, prediction and the building of smart machines. Because of the generalization principle which is based on the reduction of structural risk, SVMs are likely to capture very large areas (Nayak, Naik, & Behera, 2015).

2.6 Critical Review

2.6.1 Previous Research on Spam

The new social network has become famous for the dissemination of information that makes online users eager to publicize the latest events and personal opinions in order to reduce the user experience with huge spam posts. A study by (Qiang Zhang, Chenwei Liu, Shangru Zhong, & Kai Lei, 2017) proposed to use semantic analysis to create a self-extended dictionary which upgrades and extends dynamically with new cyber terms to spot spam comments in Chinese social media. The semantic analysis provides additional semantic functionality that aid in the identification of documents. Four text-based features that essentially represent Chinese spam commentary features have been selected using the statistical analysis of the comments of microblogging and for classifier detection, spam dictionary and text-based features is used. Comment data was gather using the Sina Weibo API where constantly rake comments from microblogs published between 12/15/2015 and 02/14/2016 by ten popular users in view of how many comments was published. An impressive 93.6 percent identification accuracy compared to conventional spam

detection approaches of experimental results indicate that the system identifies spam comments successfully in Chinese microblogging.

A research by (Kanaris, Kanaris, Houvardas, & Stamatatos, 2006) explore an alternative low-level display based on n grams, which will avoid the use of tokenizers and tools depending on language. On the basis of studies with two renowned standards, it has proven that n-grams are more accurate than word-tokens despite an increase in the aspect of the problem. It is focused on several tests of assessment. A method for extracting variable-long character n-gram based on an existing approach originally used to extract multiword conditions for data recovery applications has been proposed in addition to the character n-gram fixed length method.

In recent years a number of methods for resolving the problem of spam reviews have been suggested. A research by (Hussain, Turab Mirza, Rasool, Hussain, & Kaleem, 2019) proposed a comprehensive review of current spam screening research using the SLR methodology and validated studies focused on the features derived from study data sets and various methods and strategies used for solving the spam screening issue. This study also analyzes various metrics used for the assessment of spam detection methods. This work has established interdependencies in performance factors of any spam filtering process. The extraction of the feature is dependent on the review data set and on the selection of the functional engineering approach the accuracy of the spam detection methods. Therefore, these factors must be considered together to successfully implement the spam review detection model and to achieve better accuracy. The first systematic review of existing studies in the area of spam filtering using the SLR method.

Frustrating perspectives have drawn a lot of attention from studies as social media users have grown rapidly. The identification is still a problem given the existence of a wide range of opinion and classification techniques. A study by (Shojaee, Murad, Azman, Sharef, & Nadali, 2013) stated that the researching classifications, such as Support Vector Machine (SVM), for Sequential Minimal Optimization (SMO) and Naive Bayes, was suggested utilizing stylometric features like lexical and syntactic to spot false views. SMO uses a combination of both lexical and syntactic features to achieve the best F-measure value. The studies have confirmed

the utility of stylometric technology to identify deceit through promising results. Nonetheless, there are better alternatives and they can improve performance more, which needs more study.

Author	Tittle	Result/Description	Dataset	
Qiang Zhang	Spam	An impressive 93.6 percent	Sina Weibo	
Chenwei Liu	comments	identification accuracy	API	
Shangru	detection with	compared to conventional spam		
Zhong	self-extensible	detection approaches of		
Kai Lei	dictionary and	experimental results indicate that		
	text-based	the system identifies spam		
MA	features	comments successfully in		
A.	A.C.	Chinese microblogging.		
Kanaris,	Words Svs.	On the basis of studies with two	Ling-Spam	
Ioannis	character n-	renowned standards, it has	and	
Kanaris,	grams for anti-	proven that n-grams are more	SpamAssasin	
Konstantinos	spam filtering	accurate than word-tokens		
Houvardas,	El J. ahmul	despite an increase in the aspect		
Ioannis	oannis of the problem		_	
Stamatatos,	RSITI TEKNI	KAL MALAYSIA MELAKA	1	
Efstathios				
Hussain,	Spam Review	The extraction of the feature is	Synthetic	
Naveed	Detection	dependent on the review data set	review	
Turab Mirza,	Techniques: A	and on the selection of the	spamming	
Hamid	Systematic	functional engineering approach		
Rasool,	Literature	the accuracy of the spam		
Ghulam	Review	detection methods. Therefore,		
Hussain,		these factors must be considered		
Ibrar		together to successfully		
Kaleem,		implement the spam review		
Mohammad		detection model and to achieve		
		better accuracy		

Table 2.1: SPAM literature

Shojaee,	Detecting	SMO uses a combination of both -
Somayeh	deceptive	lexical and syntactic features to
Murad,	reviews using	achieve the best F-measure
Masrah	lexical and	value. The studies have
Azrifah Azmi	syntactic	confirmed the utility of
Azman,	features	stylometric technology to
Azreen Bin		identify deceit through
Sharef,		promising results.
Nurfadhlina		
Mohd		
Nadali,		
Samaneh		

2.6.2 Previous Research on YouTube Spam

Since the development of Internet, the classification of test also known as text categorization has become very popular. A study by (Uysal, 2018) address this question by looking at a rating system to determine the importance for each content of any post on YouTube. The experiment was conducted by gathering YouTube information and by processing YouTube data from five data sets Psy, KatyPerry, LMFAO, Eminem and Shakira. The first 100 comments for each video were first collected using the YouTube Data API and the posts that made in a different language from English were then erased. Moreover, before classifying comments using the neural network, the subject was extracted using the WordNet and Mallet library for each post. They built a system to identify spam comments via natural language processing, machine learning methods and used a benchmark dataset and on this dataset analyzed the results of three classifiers namely decision tree, Vector support and naive Bayes. For the filtering of spam comment on YouTube, the performance of five successful text feature selection methods was studied in this study. The utility of such methods was evaluated by the Naïve Bayes (NB) and the Decision Tree (DT) classifications. This study used the Macro-F1 success measure where it showed that Distinguishing Feature Selector (DFS) and Gini Index (GI) selection methods were mostly used for the highest classification results. Nonetheless, for some instances, the

choice of discriminatory features (DFSS) and the relative discriminatory criterion (RDC) tend to be less effective.

Next, a study by (Alam, 2019) present a technique for figuring out the YouTube video spam comments. It is for how to detect the comments made by those spam users who post on their own marketing purposes or to detect the comments made by those users that are not important to the provided video. Different techniques were examined for classifying spam comments with the classification of regular user comments and for the recent trend in this field as shown on figure 2.10. In conclusion, because no single method is successful in all available data sets, the cascading with best selector algorithms of various machine learning classifiers can be used to improve the result further. The message document can also be preprocessed using a natural language processing to normalize data because the words used in comment usually include slang, idioms, emoticons, symbols and abbreviations.



S.No	Research Article	Techniques Used	Available Dataset	Results
1	N-Gram Assisted YouTube Spam Comment Detection	RF,SVM, Naïve Bayes,N-Gram	13000 Comments https://developers.google.com/youtube/v3/docs/c ommentThreads#Retrieve_comments [11]	N-Gram Outclassed Other Machine Learning Algorithms
2	A Comparative Analysis of Common YouTube Comment Spam Filtering Techniques	AGA, ICA- Amuse, ELM- AE, ANN-BP, SVM-K, K-NN, LR, NBC, DT	100 Channels having 10,000 Samples https://developers.google.com/voutube/v3/ [12]	Adaptive Genetic Algorithm Performed better than other 8 Algorithms
3	A Novel Approach for YouTube Video Spam Detection using Markov Decision Process	Markov Decision Process, Decision Tree, Naïve Bayes, KNN, Random Forest, Ripper, Clustering	50 Videos, 2054 Instances out of which 824 was Spam Comment and 1230 normal	Markov Decision Process accuracy 78.52 % better than the best RF which was 72.82 %
4	Spam Detection Using KNN and Decision Tree Mechanism in Social Network	KNN, Decision Tree	FED Real Dataset http://mashable.com/2012/12/18/twitter-200- million-active-users/ Accessed July 22, 2016 [13]	Precision Call, F Measure and Class, FP and TP rate.
SULTERNIE	TubeSpam: Comment Spam Filtering on YouTube	Naïve Bayes, Decision Tree, Logistic Regression	5 Different YouTube Video Datasets Datasets / YouTube ID / # Spam/ # Ham / Total 1) Psy9bZkp7q19f0 175 175 350 2] KatyPerryCevxZvSJLk8 175 175 350 3) LMFAO KQ6zr6kCPj8 236 202 438 4) Eminemuel Hw18o7 U 245 203 448 5) Shakira pRpeEdMmmQ0 174 196 370 http://dcomp.sor.ufscar.br/talmeida/youtubespam_ collection/ [14]	A confidence level of 99 % on all algorithms, Suggested their own App. Tube Spam.
A	1 (+ 1
JNI	A Data Mining Based Spam Detection System For YouTube	Natve Bayes Decision Tree, Clustering	Self Generated Data using Tube Kit http://www.tubekit.org/ [15] *** 1. No of Videos 1719 2. No of distinct users 1428' SIA M 3. No of comments T0102865 4. No of ratings count 23013568 5. No of different categories 16	Lift. Score is Calculated. Naive Bayes and Decision Tree performed better having an accuracy of 80.20 % and 82.11 % respectively

Figure 2.10: Classifying spam comments of recent trends

A research by (Samsudin et al., 2019) proposed features to enhance the YouTube spam detection that consists of five stages, for instance data collection, preprocessing, sorting and extraction functions, classification and detection. The two forms of data mining methods have been tested and validated for each stage. The datasets used in this paper is YouTube comment from Youtube Spam collection data set, which includes five (5) comments from a total of 1956 comments from Youtube videos. Of the 1005 comments, the rest are spam. These functions are constructed using data from YouTube Spam dataset using Naïve Bayes and Logistic Regression and tested by Weka and Rapid Miner in two different data mining tools. Thirteen features evaluated at Weka and RapidMiner have been highly accurate and are therefore used in this work throughout the study. The performance of Naïve Bayes in Weka is somewhat better than the findings of RapidMiner, and the Naïve Bayes output is 87.21% and 85.29% higher than the Logical Regression in Weka. While the exactness of Naïve Bayes and its logistic recovery in RapidMiner differs slightly, 80.41% to 80.88%. But Naïve Bayes ' precision is superior to the regression in logistics. These result can be seen on figure 2.11.

Techniques	V	Veka	Rapio	lMiner
	Accuracy (%)	Precision (%)	Accuracy (%)	Precision (%)
Naïve Bayes	87.21	87.2	80.41	75.27
Logistic Regression	85.42	85.7	80.88	74.13

Figure 2.11: Results based two techniques used

When online social networks (OSN) expands, unethical efforts are rising over this form of web service. This is why many people are impacted every day by online spam and thus their confidentiality is compromised. A study by (Ezpeleta, Iturbe, Garitano, de Mendizabal, & Zurutuza, 2018) proposed a new way to verify hypothesis that current social spam filtering results can be improved by extracting the mood from the text. Different experiments with and without mood feature are done using a social spam dataset then compare the results obtained and demonstrate that mood assessment will help improve the performance of social spam filtering. Results indicate that the maximum reliability with this original data collection is enhanced by a Youtube Comments Dataset from 82,5% to 82,58% and by a validation dataset, from 93,97% to 94,38% on figure 2.12. In addition, on average 13.76% and 11.47% are reduced to false positive as shown in figure 2.13.

	Ne	ormal	M	lood	-	
Name		Acc	FP	Acc		FP reduction (%)
NBM.c.stwv.go.ngtok	89	82.50	77	82.53	-	13.48
NBMU.c.stwv.go.ngtok	89	82.50	70	82.58		21.35
NBM.stwv.go.ngtok	71	82.48	63	82.43		11.27
NBMU.stwv.go.ngtok	71	82.48	59	82.43		16.90
NBM.c.stwv.go.ngtok.stemmer	81	82.45	73	82.45		9.88
NBMU.c.stwv.go.ngtok.stemmer	81	82.45	68	82.48		16.05
NBM.stwv.go.ngtok.stemmer	64	82.35	58	82.38		9.38
NBMU.stwv.go.ngtok.stemmer	64	82.35	53	82.35		17.19
CNB.stwv.go.ngtok	125	82.30	110	82.43		12.00
CNB.stwv.go.ngtok.stemmer	109	82.28	98	82.20		10.09
					avg:	13.76

Figure 2.12: First data set

No	ormal	Λ	lood	
\mathbf{FP}	Acc	\mathbf{FP}	Acc	FP reduction (%)
85	93.97	76	94.38	10.59
89	93.87	85	94.02	4.49
113	92.69	80	94.17	29.20
113	92.69	116	92.54	-2.65
119	92.38	86	93.97	27.73
119	92.38	123	92.23	-3.36
127	92.13	96	93.66	24.41
127	92.13	127	92.13	0.00
135	91.72	101	93.35	25.19
135	91.72	137	91.62	-1.48
	2			avg: 11.41
	Na FP 85 89 113 119 127 127 135 135 135	Normal FP Acc 85 93.97 89 93.87 113 92.69 113 92.69 119 92.38 119 92.38 127 92.13 127 92.13 135 91.72 135 91.72	Normal M FP Acc FP 85 93.97 76 89 93.87 85 113 92.69 80 113 92.69 116 119 92.38 86 119 92.38 123 127 92.13 96 127 92.13 127 135 91.72 101 135 91.72 137	Normal Mood FP Acc FP Acc 85 93.97 76 94.38 89 93.87 85 94.02 113 92.69 80 94.17 113 92.69 116 92.54 119 92.38 86 93.97 119 92.38 123 92.23 127 92.13 96 93.66 127 92.13 127 92.13 135 91.72 101 93.35 135 91.72 137 91.62

Figure 2.13: Validation data set

A study by (Aiyar & Shetty, 2018) proposed on the YouTube video streaming platform a technique to track unwanted feedback or spam and have shown that the use of the character-gram (n-character substrates) over words-grams increases the classification accuracy. The dataset compiled nearly 13,000 commentary manually using the open Youtube API from different Youtube channels, and processed it in a server for evaluation. We see that the character grams are better able than word grams to identify the measurement of spam ability in the comment. In addition, the Vector Machines & Random Forest are more conventional and more adaptable to highdimensional datasets than other mainstream machine learning algorithms.

Author	Tittle	Result/Description	Dataset
Uysal, Alper	Feature	Naïve Bayes (NB), Decision Tree (DT)	Daily
Kursat	Selection	classifications and Macro-F1 success	Telegraph
	for	measure was used where	and
	Comment	Distinguishing Feature Selector (DFS)	YouTube.
	Spam	and Gini Index (GI) get highest	
	Filtering on	classification results, discriminatory	
	YouTube	features (DFSS) and relative	
		discriminatory criterion (RDC) tend to	
		be less effective.	
Alam,	Spammer	No single method is successful in all	-
Rafaqat	Detection:	available data sets, the cascading with	
A. C.	A Study of	best selector algorithms of various	
EKN	Spam Filter	machine learning classifiers can be	
TU	Comments	used to improve the result further.	
1000	on YouTube		
-91h	Videos		
Samsudin,	Youtube	The performance of Naïve Bayes in	YouTube
Nur'Ain	spam	Weka is somewhat better than the	Spam
Maulat NIVE	detection	findings of RapidMiner, and the Naïve	Collection
Mohd Foozy,	framework	Bayes output is 87.21% and 85.29%	Data Set
Cik Feresa	using naïve	higher than the Logical Regression in	from
Binti	bayes and	Weka. While the exactness of Naïve	Machine
Alias,	logistic	Bayes and its logistic recovery in	Learning
Nabilah	regression	RapidMiner differs slightly, 80.41% to	Repository
Shamala,		80.88%. But Naïve Bayes ' precision is	
Palaniappan		superior to the regression in logistics	
Aiyar,	N-Gram	We see that the character grams are	Open
Shreyas Shetty, Nisha	Assisted	better able than word grams to identify	Youtube
P.	Youtube	the measurement of spam ability in the	API from
	Spam	comment. In addition, the Vector	different
		Machines & Random Forest are more	

Table 2.2:	YouTube	SPAM	literature
-------------------	---------	-------------	------------

Comm	ent conventional	and more adapta	able to Youtube
Detect	ion high-dimens	ional datasets than	n other channels
	mainstream	machine le	earning
	algorithms		

2.7 Conclusion

This section provides a better description of internet security attack, classification of attack, spam machine learning and critical review. Every topic is generally described to illustrate how these approaches can be applied in this task. A hypothetical analysis of previous research into spam and YouTube spam was suggested in the critical review section. The findings from the analysis are used to establish an active study system in the detection of Internet security attacks. The chapter will explained in details on the methodology which greatly influenced with this chapter's findings



CHAPTER 3: PROJECT METHODOLOGY

3.1 Introduction

This chapter provides a detailed description of the methodology used as a guide for ensuring a successful project is always in progress. The approach also includes guidelines or protocols for certain benchmarks in the research process. This is essential to ensure that the plan is successfully implemented within a specific timeframe. A Gantt graph was chosen to show the timeline of this plan and the objectives. The flux maps are also used to define the series process and associated studies of this project in its entirety.

3.2 Methodology I TEKNIKAL MALAYSIA MELAKA

Methodology is a collection of methods or approaches to address specific issues by some strategies. The technically analytical approaches that act as a context explaining how the system meets task objectives is used in this research methodology. Six phases involve previous research, information gathering, definition of scope, design and implementation, system testing and assessment and lastly documentation into the process of this Project methodology. The relationship of each phase is described as the unique methodology model used in this project in Figure 3.1.



This stage offers an increased view into the execution of the project on the basis of previous studies. The process will assess domain articles based on their limitations of research and future research recommendations so that an important research topic can be identified. The field studied in this project is internet security attack, machine learning and classification. Previous research will thus provide an overview of how the planned theoretical frameworks operate in their respective fields. This is discussed in detail in the phase.

3.2.2 Information Gathering

The collection of information is intended to develop a deep understanding of research issues. The proof of how critical the issues are is verified. This work, on the other hand, uses SVM classification system. The collection of information from previous research will reinforce the selection of algorithms and methods for this research.

3.2.3 Define Scope

The scope of the project is the limit of the area of this study. This research conducted to analysis the data set of Shakira obtained from UCI Machine Learning Repository. The method that looked into is Machine Learning method only.

3.2.4 Design and Implementation

This section details the design and implementation of the new models on the basis of the previous studies. First, the dataset is obtained from UCI Machine Learning Repository. Next, data set is separated into a few considerate features. Followed by the data set is extracted using lexical features to be trained. The new set of features are then used in classification using SVM which later produced the best features model to be considered.

3.2.5 Testing and Evaluation of Model

The evaluation and testing phase was designed to determine whether the model developed met the prescribed requirements. In this case, the features developed at the end of the process will be compared its accuracy. However, it was predicted that a lexical analysis by SVM would give a better accuracy to literature from previous research. This work aims therefore to test this theory with multiple studies. Future work on development of this research could be established by comprehensive design analysis.

3.2.6 Documentation

The process of documentation helps to structure the results in a more structured and appropriate manner. All processes, constraints and results in a correct documentation, acting as reference and evidence of every activity, will be written correctly for each experiment. To order to achieve its aims, this project requires certain complicated procedures. It will be a burden without a good reporting strategy to keep track and handle it all. In the respective sections, all relevant information is first identified and documented.

3.3 **Project Schedule and Milestones**

The project schedule of the plan determines the tasks or processes involved. The milestones are set at certain phases in order to always keep the project on track. A Gantt diagram is used for this purpose to describe timetables and milestones. A flux map also provides a systematic overview of the tasks and their relationship. This stage will also identify the resources needed in order to avoid delays or future constraints. In order to complete this project within a time-frame, each task is given a specific completion period. Excellent project management and scheduling contribute to a very good result.

3.3.1 Project Flowchart

Figure 3.2 shows the overall project phase flowchart.

Figure 3.2: Project Flowchart

3.3.2 Project Milestones

The timeframe for this project is shown in Table 3.1

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

		(28/10/2019) - (1/11/2019)	MID
			SEMESTER
			BREAK
		(4/11/2019) - (8/11/2019)	Chapter 4:
		Information collection and	Analysis and
		analysis.	Design
10-15	Design	(11/11/2019) – (15/11/2019)	Chapter 4:
		Design the project and choose the	Analysis and
		tools for implement	Design
		(18/11/2019) – (22/11/2019)	Progress report
		Design the environment for	on Chapter 4
	LAYSI	implementation.	
AT IN	40	(25/11/2019) – (29/11/2019)	PSM1 Report
No.		Write and finalize project report.	
T T	-	(2/12/2019) - (6/12/2019)	PSM1 Report
FIGURA		Submit project report to supervisor.	
de l	((9/12/2019) – (13/12/2019)	Presentation
ملافح	مليسيا	Schedule the Presentation	Schedule
UNIVE	RSITI TE	KNIKAL MALAYSIA MELA	KA

Table 3.1: Milestone

3.3.3 Project Gantt chart

Refer to appendix section.

3.4 Conclusion

In brief, strategy plays an important role in determining a project's progress. This helps to define the entire process required to achieve the goals. The phases and approaches of each of them are discussed in this chapter. The main goal of this project is to measure the accuracy of spam comments on Youtube by all approaches and methods. The subsequent section will explain more about the methods and techniques presented here. The next section will discuss in detail the method and steps of the design a lexical features of learning machines using SVM.

CHAPTER 4: ANALYSIS AND DESIGN

4.1 Introduction

The problem statements are discussed further in this section, where the reasonable approach is specified for each step of the previous chapter. All the processes concerned are identified and analyzed first before moving to the next phase in order to ensure that these research achieve the aims. The best method to use in this experiment is the Chi-Square and Information gain method for the feature selection. All approaches have been selected given its established literary quality. The features are then fed into a series of experiments with several approaches to machine learning algorithms. In the effort to determine the best configuration for this work, software and hardware specifications are also evaluated.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

The design phase includes a further analysis of each of the methods used for the experiments. The process of experiments is illustrated and their relationship with each other is indicated by the structure of projects design. It reveals some unseen limitations that could interrupt the phase of development. The analysis and design stages also leads to a clear understanding of the nature of the plan and a detailed model to prevent possible potential mistakes.

4.2 Problem Analysis

The main goal of this research is develop a machine learning model in which ham and spam comment can be differentiated. The detection process is done by the pre-program learning algorithm automatically. The concern is that the data set use contains five features which are comment_id, author, date, content and class.
Nonetheless, this analysis focuses only on the two last features which are content and class features since it is difficult to distinguish between spam and ham comment in order to cover all five features. A large number of instances require high computing resources and a high-precision machine learning algorithm. YouTube Spam Classification will be solved by this study. The method of classification is the Support Vector Machine (SVM), which is commonly used in machine learning. The best method of solving problems is this approach.



4.3 Project Design

Figure 4.1: Project Design

Figure 4.1 illustrates the overall concept for the spam comment identification machine learning algorithm. The process begins with acquiring dataset, followed by preprocessing, feature extraction (identify features descriptors) and generate BOW feature vector (TF). Next are splitting train and test sets which is divided into two process. First directly go into normalization, classification (SVM), Bag of word model and accuracy for comparison. Second is feature selection that consist information gain and chi square. Later on, it divided into two separate process generate information gain feature vector (TF), normalization, classification (SVM), information gain model and accuracy for comparison. Next, generate chi square feature vector (TF), normalization,

classification (SVM), chi square model and accuracy for comparison. Everything in this section will be clarified.

4.3.1 Dataset

The datasets used in this paper will be YouTube comment from the Machine Learning Repository's YouTube Spam Collection Data Set. Five YouTube video comments with a total of 1956 comments were obtained in the datasets. A total of 1005 comments are spam, while the rest are comments from ham. Table 4.1 indicates the focused datasets. According to (Samsudin et al., 2019) this dataset is also used in the conduct of their study. In fact, these datasets are a publicly available database on the Web.

3	Dataset		Spam	ham	Total
TEKW	Psy	N.K.A	175	175	350
ALSON D	Katy Perry		175	175	350
K	Eminem	کل	245	رسيني ني د	⁴⁴⁸ .
	LMFAO /ERSITI TE	EKN	236 IKAL MA	LAYSIA ME	438 ELAKA
	Shakira		174	196	370
ТО	/TAL		1005	951	1956

Table 4.1: Description of Dataset.

A	В	C	D	E
COMMENT_ID	AUTHOR	DATE	CONTENT	CLASS
z12pgdhovmrkt	lekanaVE	2014-0	i love this so much. AND also I Generate Free Leads on Auto Pilot & amp; You Can Too! http://www.MyLeaderGate.com/moretraffici» ¿	1
z13yx345uxepet	Pyunghee	2014-0	.http://www.billboard.com/articles/columns/pop-shop/6174122/fan-army-face-off-round-3 Vote for SONES pleasewe're against vipsplease help us & gt; & & & & & & & & & & & & & & & & & & &	. 1
z12lsjvi3wa5x1v	Erica Ross	2014-0	Hey guys! Please join me in my fight to help abused/mistreated animals! All fund will go to helping pay for vet bills/and or helping them find homes! I will place an	1

Figure 4.2: Example of Kate Perry dataset

Figure 4.3: Example of features will be use

4.3.2 Data Preprocessing

The process of purging and text preparation for classification is the preprocessing of the data. Online texts typically contain a lot of noise and data, such as HTML tags, scripts and advertisements. The fact that these terms exist makes it more difficult to understand the nature of the question and thus the identification since each term in the document is viewed as a single aspect. The idea is that the details are properly prepared in order to reduce noise in the message, which should help improve the identification quality and thus speed up the classification method, thus helping to interpret real-time sentiment analysis (Haddi, Liu, & Shi, 2013). It senses specific characteristics and deletes irrelevant, obsolete and noisy information. This method speeds up algorithms for data mining, boosts statistical precision and enhances understandability (V. Kumar, 2014). This study divided data processing into six sub processors namely tokenization, token length , case normalization, special character, stop word removal and stemming.

i love this so much. AND also I Generate Free Leads on Auto Pilot & amp; You Can Too! http://www.MyLeaderGate.com/moretraffici'*; http://www.billboard.com/articles/columns/pop-shop/6174122/fan-army-face-off-round-3 Vote for SONES please....we're against vips....please help us.. >.<i'*; Hey guys! Please join me in my fight to help abused/mistreated animals! All fund will go to helping pay for vet bills/and or helping them find homes! I will place an extra emphasis on helping disabled animals, ones otherwise would just be put to sleep by other animal organizations. Donate please. http://www.gofundme.com/Angels-n-Wingzi'*;

Figure 4.4: Raw data

The method of tokenization includes collecting and breaking down the contents of the comments. Here the phrases are separated by white spaces, symbols, or special characters, into words and tokens. Characters such as the basic delimiter, brackets, mathematical operators and special characters contained in the word will also be diminished. i love this so much . AND also I Generate Free Leads on Auto Pilot & amp ; You Can Too ! http : //www.MyLeaderGate.com/moretraffic http : //www.billboard.com/articles/columns/pop-shop/6174122/fan-army-face-off-round-3 Vote for SONES pleasewe 're against vipsplease help us.. & gt ; . & It ; Hey guys ! Please join me in my fight to help abused/mistreated animals ! All fund will go to helping pay for vet bills/and or helping them find homes ! I will place an extra emphasis on helping disabled animals , ones otherwise would just be put to sleep by other animal organizations . Donate please . http : //www.gofundme.com/Angels-n-Wingz

Figure 4.5: Tokenization

However, every word extracted from the tokenization process needs to be refined. Thus, the process of token length implemented by the number of letters per word processed is between 2 and 10. Words with letters of 1 and more than 10 will be recorded only in frequency.

AALAYSIA i love this so much . AND also I Generate Free Leads on Auto Pilot & amp ; You Can Too ! http Vote for SONES pleasewe 're against vipsplease help us.. & gt ; . & It; Hey guys ! Please join me in my fight to help animals ! All fund will go to helping pay for vet bills/and or helping them find homes ! I will place an extra emphasis on helping disabled animals , ones otherwise would just be put to sleep by other animal organizations. Donate please . Figure 4.6: Token Length

Followed by that, a case normalization will be used to modify the entire comment in the lower or upper case words. Lexical analysis is capable of extracting comments into words and removing special characters and converting all letters of a word to lowercase letters.

i love this so much . and also i generate free leads on auto pilot & amp ; you can too ! http vote for sones pleasewe 're against vipsplease help us.. & gt ; . & It ; hey guys ! please join me in my fight to help animals ! all fund will go to helping pay for vet bills/and or helping them find homes ! i will place an extra emphasis on helping disabled animals , ones otherwise would just be put to sleep by other animal organizations . donate please .

Figure 4.7: Case Normalization

Spammers use a common technique to avoid spam filters to misinterpret the message they are sending in the subject lines (Krishnamurthy, 2015). It is often

observed that numeric and special features (1, 2, 5, etc.) and special features (@, number, percent, etc.) have no effect on the analysis. Nonetheless, during transformations of text files to numeric vectors they also cause confusion (Tripathy et al., 2016). Thus, by removing special character helps the preprocessing process to be processes in much better.

i love this so much and also i generate free leads on auto pilot amp you can too http vote for sones please we re against vips please help us gt It hey guys please join me in my fight to help animals all fund will go to helping pay for vet billsand or helping them find homes i will place an extra emphasis on helping disabled animals ones otherwise would just be put to sleep by other animal organizations donate please

Figure 4.8: Special Character

Stop word removal can reduce the size of the database while allowing only selected words to be processed. The statement text generally contains unnecessary words, such as 'was', 'the', and 'a'. Such terms are not helpful in the analysis of spam comments so it is better to remove them to avoid noise and unnecessary tokens. Take a review for instance, "It's an excellent car." The review appears as "good car" after removal of stop words and punctuations.



Figure 4.9: Stop word removal

However, there are also words that have different spelling but have the same meaning or are simply referred to as having the same root word. Since stop words were very common words, which do not contain any information (such as pronouns, prepositions, conjunctions etc.) it is possible to improve cluster results by removing stop words. To address this problem, a process known as stemming must be implemented on words.

A stemming algorithm converts the various word shapes into one recognized form. For example the terms "works", "working", and "worked" as instances of the word work. It reduces the tokens or words to its root form this means the removal of suffixes, which produces words that have the same conceptual meaning. If all these words were used as input attributes, then the database would be large. This method reduces the number of characteristics in the space vector and improves the rate and identification of training for many classifier. Yet stemming may contribute to the stemming of two different words as the same phrase.

love gener free lead auto pilot vote sone pleas re vip pleas help us gt It guy pleas join fight help anim fund go help pay vet billsand help home place extra emphasi help disabl anim one otherwis anim organ donat pleas

Figure 4.10: Stemming

For Table 4.2 indicates the advantage and disadvantage implementing the rules for this study.

WALAYS/4

Rules	Advantage	Disadvantage			
EK					
Lexical	Breaking down the contents of	Token size will be too large then			
Analysis	the comments by diminished	hard to distingue between the			
aunt.	unusable character. Process	usable outputs.			
املاك	only selected range number of	او بیقہ سینڈ ا			
	letters per word.	Q. 0			
UNIVER	RSITI TEKNIKAL MALA	YSIA MELAKA			
Stop Word	Improve clustering results.	Words will increase the size of			
Removal	Reduce the size of the	the vector space thus			
	database while allowing only	complicating the categorization			
	selected words to be	process.			
	processed.				
Stemming	Transform words to their	Stemming may cause two			
	roots. Reduces the number of	different words to be stemmed as			
	features in the space vector	a same word.			
	and increases the learning				

	speed and categorization	
	phases for many classifiers.	
Case	Changes the whole comment	The comment that have same
Case	changes the whole comment	The comment that have same
Normalization.	either in lower case letters or	meaning will be redundant.
	upper case letters.	

Table 4.2: Advantage and disadvantage of rules

4.3.3 Feature Extraction

As for feature extraction, this research using word frequencies. The last process of data preprocessing will be used to generate a feature descriptor as in figure 4.11.

love gener free lead auto pilot vote sone pleas re vip help us gt It guy join fight anim fund go pay vet billsand home place extra emphasi disabl one otherwis organ donat **Figure 4.11: Feature Descriptor** Finally, the feature descriptor will be generated as in figure 4.12. love gener free lead auto pilot vote sone pleas re vip help us gt it guy join fight anim fund go pay vet billsand home place extra emphasi disabl one otherwis organ donat 000000.111111110000000000000000000000 0000000000000011111111111111111111111

Figure 4.12: Feature Vector

4.3.4 Generate BOW feature vector

Word bag (BOW) was produced to test its reliability by contrasting it with the new model established at the conclusion of these studies. Datasets must gather from UCI's Machine Learning Repository and split the dataset into two sections, the first part for train and the second part for testing, until BOW is produced. The normalization for train and test dataset is performed after that next step. Normalization is a data preprocessing procedure that is significantly affected by the function and the variable is considered to be eliminated.

4.3.5 Data Splitting and Validation

Text can be tokenized using term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TFIDF). For this study, the implemented method are only using TF. To do this, the YouTube Spam comment dataset data are divided equally subsampling as figure below.



As the dataset is finite it can be processed by two approaches. First used the entire training data to select the best classifier and determine the error rate. This problem is linked to two fundamental issues that are likely to over fit the training data, and the error rate is too optimistic or below the true error rate. So to overcome it, the second solution for this project was developed and used. The approach is to split the training data into disjoint subsets using random subsampling. A set number of instances without substitution are selected randomly for each group.

In this case, the total number of instances from both ham and spam data are 350. Then, the random subsampling technique are used to generate 160 random instances for train and test data which each of it contain 80 ham and 80 spam instances respectively. The train and test data generated with balance number of instances to

fairly train the classification algorithm to detect ham and spam data, thus producing fair result.

4.3.6 Feature Selection

During the feature selection process, train set information features are evaluated with Chi-Square and Information Gain to verify their output to assess predictive precision. The main objective of the choice of functions is to eliminate nonimperative functions from the list to reduce the machine complexity.

In Chi-Square the target variable is tested to depend on the number data function variable. If the feature variable is independent, it is regarded as not significant and discarded.

The target variable class label is evaluated using chi-quare statistics and observe the relationship between target and variable feature. If the class label has an independent relation with the class factor, it is discarded. On another hand, if they are dependent, the feature variable considered as very important which is closely related to the result/outcome. To evaluate whether the feature variable and class label are dependent a certain threshold is used. A null hypothesis and a corresponding alternative hypothesis of chi-square method as below.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

H0: The two variables are independent

H1: The two variables are related.

The H0 (null hypothesis) is discarded when the p-value obtained from the experiment is less than the per-detailed significance level (usually of 0.005).

The formula for Chi-square feature selection (Wuensch, 2011) are as follow

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Oi is the category incidence and Ei is the class presence in the above equation. Based on these two factors, a rank is calculated. The importance of the attribute for that dataset is shown by the chi-squared algorithm running on a dataset. The number of main attributes needed for predicting the class variable is selected based on the attribute importance. Two forms for Chi-square analysis. The fit test Chi-square decides whether the survey information correlates to the group. In the FSelector library R, Chi-square is present as a variable.

The selection of information gain is rendered in Weka. Weka is a collection of data mining machine learning algorithms. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

The Ranker Search System uses the InfoGainAttributeEval attribute evaluator to evaluate the data gain (entropy) for each outgoing parameter attribute. Output values vary between 0 (no information) and 1 (maximum data). The more insightful qualities would possibly get a greater value and be picked, while the less will have a lower score and will be discarded.

The measurement of information gain is based on the concept of entropy. It is often used as an indicator of features relevance in filter techniques that determine individual characteristics and the benefit of this method is that it is fast (Alhaj, Siraj, Zainal, Elshoush, & Elhaj, 2016). Let D (A1, A2... An, C), n x 1, be an array of n+1based information with C to be a class-based attribute. Let m be the number of separate values of class.

In the next step, all selected features and the information gain are used in contrast to which is the best way to achieve better prediction precision in the study.

4.3.7 Normalization

Normalization is one of the technique involved in data preparation for machine learning. It changes the values of numeric columns in the dataset to use a more understandable and common scale without affecting the range of values. It transform the dataset into a more understandable form. This research will normalize the data to a value ranging from 0 to 1.

4.3.8 Classification

SVM is a supervised classification or regression machine learning algorithm. SVM is used for classification tasks in this study. The value of the given coordinate as in Figure 4.14 shows SVM plotting each data item in n-dimensional space (n= Number of features) with each value.



Now, to draw a decision boundary (hyperplane) in order to distinguish between two classes of SVM, to be separated or classified as in Figure 4.4. There are several parameters that are provided to determine the best place to draw hyperplane In order to separate the groups, the plane must be drawn as close as possible. The nearest data point to the hyperplane is defined as the support vectors of the form of the SVM. The most ideal hyperplane, which is the shortest from support vectors, and the distance between support vectors and hyperplanes is known as the margin. The hyperplane with support vectors (green circle) and margin illustration shows in Figure 4.15.



Figure 4.15: Hyperplane Placement



Figure 4.16: Hyperplane with Margin

The right position can be challenging to choose for the hyperplane. Figure 4.16 indicates that line B has a higher margin in comparison to A, the strongest hyperplane to always choose the largest margin in accordance with the basic concept of SVM. Hyperplane B does however have a classification failure and A correctly classified everything. Although the margin is smaller, hyperplanes A are more reliable than B. The formula for the accuracy calculation of a hyperplane as shown in Table 4.17.



Figure 4.17: Choosing the Best Hyperplane

The following equation is used to create confusion for the above image,



False Negative Ratio (FNR)

Where

True Positive: the no. of spam correctly detected.

True Negative: the no. of ham correctly detected.

False Positive: the no. of ham identified as spam

False Negative: the no. of spam identified as ham.

Next, the following confusion matrix can be made:

Blue vector for ham,

Red for spam

Hyperplane	TP	FN	FP	TN	Accuracy
А	6	0	0	6	100%
В	8 ALAYSIA	0	1	5	72%

Table 4.3: Confusion Matrix of Hyperplane A and B

In some instances, as in Figure 4.18, one of the details can be located in the other area called outlier.



Figure 4.18: Graph with an Outlier

SVM has a feature to ignore outliers and find the best hyperplane with the maximum margin to deal with this problem. This is one of the advantages of SVM for outliers where it is robust.

Above situation is called the Liner SVM, with a straight line that divides vectors. The majority of problems in real life are usually non-linear, as is shown in Figure 4.19.



Figure 4.19: Non-Linear SVM's Case

The kernel features are used in the SVM to resolve this. Kernel will aid in transforming non-linear spaces into linear space. Basically, the two variable x and y are transformed into a new space, z. As shown in Figure 4.20, the space in two dimensions is transformed into a multidimensional space. Next, this non-linear data can be separated from the hyperplane.

UNIVERSITI TEKNIKAL MALAYSIA MELAKA



Figure 4.20: SVM with Multi-Dimensional Space

An important part of a kernel is to decide how hyperplane is constructed. Table 4.4 is the analysis of the SVM kernels as in (Chih-Wei Hsu, Chih-Chung Chang, 2000).

Kernel	Description
Linear	Separates the vectors linearly by using
	straight line. Good at classifying two
	classes at a time or binary-class problem
Polynomial	Works well with non-linearly separable
	data.
Radial Basis Function (RBF)	Used to classify non-linear data. Can be
HALAYSIA MELPIXA	used for time series prediction, function approximation, classification and system control
Sigmoid	Originate from neural networks, a
AUNU	classification approach with less time
كنيكل مليسيا ملاك	consumption
UNIVERSITI TEKNIKAL M	IALAYSIA MELAKA

Table 4.4: SVM Kernel

All of them are first evaluated and the precision of the hyperplane is measured in order to choose the right kernel. The high accuracy of this plan is chosen and used.

The input vectors of nonlinear information are mapped with the correct kernel function k(i, j into a broad parameter space H). The RBF kernel is used mathematically for this plan (Vapnik 1995).

$$k(\vec{h}_i, \vec{h}_j) = \exp\left(\frac{-\|\vec{h}_i - \vec{h}_j\|^2}{2\sigma^2}\right)$$

The classification function has the following kernel format:

$$f(\vec{h}) = \operatorname{sgn}\left[\sum_{i=1}^{l} \alpha_i y_i k(\vec{h}_i, \vec{h}) + b\right]$$

Where the labeled class yi $\{+1,-1\}$ is indicated, k is a function of the kernel (RBF), b is a prejudice and I the multiplier coefficient Lagrange obtained by resolving the quadratic programming problem. Only those I with the appropriate coefficients I > 0 in the previous formula are called support vectors (SVs), which are useful for deciding on a classification problem based on Karush-Kuhn-Tucker (KKT) quadratic programming conditions. Finding coefficients i is equivalent to maximizing the function as follows:



Where C is a non-negative parameter for regularization. In some data sets, a separating hyperplane in the functional space might not be found by the SVM. A regularization variable C was therefore used to monitor the balance between optimizing the cap and reducing the error in the learning. To absolutely define a support vector machine, the kernel function and the magnitude C for the breaking of the weak margin are expected to be two parameters. Such parameters will be chosen depending on the particular model data. Grid-Search is employed for optimum hyper parameters by changing the C and Gamma function of a system, which contributes to predictors that are more reliable.

The data were normalized as a form of model predictions when the test data were carried out. The model form in training will be used as forecast while data tested will be used for YouTube spam comment detection. The forecast will be used as a comparison to other models when the second model is available. New model, applied when process detection happened through classification of accuracy. As a result, the precision of the results of this model will be compared to other accuracy produced by using BOW, CS and IG to ensure that this model detects YouTube spam and ham comment more reliable.

4.4 Requirement Analysis

In order to carry out this research successfully, many criteria were created. Such criteria are listed as software and hardware that are explained in the following sections in depth.

4.4.1 Software Requirement

The software collection used in the project as in Table 4.5.

Software	Descriptions
	اوتور شنی ت
Eclipse IDE for Enterprise Java	An Integrated Environment for
Developers. ERSITI TEKNIKAL	Development (IDE) for machine
	learning algorithm implementation in
	Java.
Windows 10 Pro	Environment of the operating system
Weka	A data training platform
Microsoft Project	Application for Gantt map design
Notepad++	Software for reading and combining data

Microsoft Word	Software used for work statement
	planning and reporting.
Microsoft Excel	Software used to sort data according to
	their attributes and instances and arrange
	it accordingly.
draw.io Desktop	Software for building flowcharts like
	diagrams.

Table 4.5: Software Requirement

4.4.2 Hardware Requirement

A Dell Inspiron 5421 is used to execute all functions from reporting to studies as a workstation. The laptop's features are listed in Table 4.6.

Specification 40	Description
Jalle	Lundo 15:5: in in aniel
Processor	Intel(R) Core(TM) i5-3337U CPU @ 1.80GHz (4 CPUs),
UNIVER	S18GHZKNIKAL MALAYSIA MELAKA
Operating	Windows 10 Pro 64-bit (10.0, Build 18363)
System	(18362.19h1_release.190318-1202)
Graphics	NVIDIA GeForce GT 730M (2GB DDR3L)
Webcam	Native HD 1.0MP webcam with digital microphone
Memory	8 GB DDR3
Storage	500 GB HDD and 240 GB SSD

Audio	Stereo speakers with Waves MaxxAudio® 4 processing Built-in
	digital microphone
Battery	6-cell/3.0Ah (65 WHr) Lithium Ion
Display	LED Touchscreen and 1366 x 768 Pixels
WLAN	Intel® Centrino® Wireless-N 2230, 802.11b/g/n + BT4.0
LAN	10/100/1000M Gigabit Ethernet
Bluetooth®	4.0
Ports	(2) USB 3.0 + (1) USB 2.0 ,RJ45 Ethernet ,HDMI TM v1.4a
s.Ph	,Combination headphone/microphone jack, Multi-media Card
ling 1	Reader ,Digital (SD) Memory Card
F	
Dimensions	346 x 245 x 32.9 mm
* JAINO	
Joll	unde Kai Gi Sur aniel

Table 4.6: Hardware Requirement

4.5 Conclusion

In short, this section describes in detail the outcomes of this analysis and all its methods. The analysis process helps to identify and decompose the modules of the research methodology. This is important to make sure that all planned work is effective in achieving the aims of research. The design phase turns the study into a set of methods for achieving the aims. All of this is done through an examination of research literature which acts as a guide to how the goals can be achieved. A model is developed using SVM which is compared with IG and Chi-Square feature selection methods for its detection accuracy.

REFERENCES

- Abdullah, A. O., Ali, M. A., Karabatak, M., & Sengur, A. (2018). A comparative analysis of common YouTube comment spam filtering techniques. 6th International Symposium on Digital Forensic and Security, ISDFS 2018 -Proceeding, 2018-Janua, 1–5. https://doi.org/10.1109/ISDFS.2018.8355315
- Aiyar, S., & Shetty, N. P. (2018). N-Gram Assisted Youtube Spam Comment Detection. *Procedia Computer Science*, 132(Iccids), 174–182. https://doi.org/10.1016/j.procs.2018.05.181
- Alam, R. (2019). Spammer Detection: A Study of Spam Filter Comments on YouTube Videos. (May).
- Alberto, T. C., Lochter, J. V., & Almeida, T. A. (2015). TubeSpam: Comment Spam Filtering on YouTube. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), (January), 138–143. https://doi.org/10.1109/ICMLA.2015.37
- Alhaj, T. A., Siraj, M. M., Zainal, A., Elshoush, H. T., & Elhaj, F. (2016). Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation. *PLOS ONE*, *11*(11), e0166017. https://doi.org/10.1371/journal.pone.0166017
- Ali, M. M. (2006). Intrusion Detection, Denial of Service (DoS) Denial of Service (DoS).
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. Retrieved from http://arxiv.org/abs/1707.02919
- Annadatha, A., & Stamp, M. (2018). Image spam analysis and detection. Journal of Computer Virology and Hacking Techniques, 14(1), 39–52. https://doi.org/10.1007/s11416-016-0287-x
- Attar, A., Rad, R. M., & Atani, R. E. (2013). A survey of image spamming and filtering techniques. Artificial Intelligence Review, 40(1), 71–105. https://doi.org/10.1007/s10462-011-9280-4
- Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. (2008). Web Spam Detection: Link-based and Content-based Techniques. *The European Integrated Project Dynamically Evolving Large Scale Information Systems DELIS Proceedings of the Final Workshop*, 222, 99–113. Retrieved from

http://www.chato.cl/papers/becchetti_2008_link_spam_techniques.pdf

- Biggio, B., Fumera, G., Pillai, I., & Roli, F. (2008). Improving image spam filtering using image text features. *5th Conference on Email and Anti-Spam, CEAS 2008*.
- Brooks, D. J. (2010). What is security: Definition through knowledge categorization. *Security Journal*, 23(3), 225–239. https://doi.org/10.1057/sj.2008.18
- Castillo, C., Donato, D., Gionis, A., Murdock, V., & Silvestri, F. (2007). Know your Neighbors: Web Spam Detection using the Web Topology. *Proceedings of the* 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07, 423. https://doi.org/10.1145/1277741.1277814
- Cateni, S., Colla, V., & Vannucci, M. (2014). A Hybrid Feature Selection Method for Classification Purposes. 2014 European Modelling Symposium, (October 2019), 39–44. https://doi.org/10.1109/EMS.2014.44
- Chavda, A., Potika, K., Troia, F. Di, & Stamp, M. (2018). Support Vector Machines for Image Spam Analysis. Proceedings of the 15th International Joint Conference on E-Business and Telecommunications, 1(Icete), 597–607. https://doi.org/10.5220/0006921405970607
- Chih-Wei Hsu, Chih-Chung Chang, and C.-J. L. (2000). A Practical Guide to Support Vector Classification. *Theory, Culture & Society, 17*(1), 39–61. https://doi.org/10.1177/02632760022050997
- Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1), 23. https://doi.org/10.1186/s40537-015-0029-9
- Dakpa, T., & Augustine, P. (2017). Study of Phishing Attacks and Preventions. International Journal of Computer Applications, 163(2), 5–8. https://doi.org/10.5120/ijca2017913461
- Danhieux, P. (2013). Email Phishing Attacks. *SANS Securing the Human*, (February), 2–4. Retrieved from http://www.securingthehuman.org/newsletters/ouch/issues/OUCH-201302_en.pdf
- Dinakaran S, & Dr. P. Ranjit Jeba Thangaiah. (2013). Role of Attribute Selection in Classification Algorithms. *International Journal of Scientific & Engineering Research*, 4(6), 67–71. https://doi.org/June 2013
- Dredze, M., Gevaryahu, R., & Elias-Bachrach, A. (2007). Learning fast classifiers for

image spam. 4th Conference on Email and Anti-Spam, CEAS 2007.

- Elleithy, K., & Blagovic, D. (2006). Denial of Service Attack Techniques: Analysis,
 Implementation and Comparison. *Journal of Systemics*, ..., 3(1), 66–71.
 Retrieved from http://www.iiisci.org/Journal/CV\$/sci/pdfs/P129065.pdf
- Ezpeleta, E., Iturbe, M., Garitano, I., de Mendizabal, I. V., & Zurutuza, U. (2018). A Mood Analysis on Youtube Comments and a Method for Improved Social Spam Detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*): Vol. 10870 LNAI (pp. 514–525). https://doi.org/10.1007/978-3-319-92639-1_43
- FireEye Inc. (2016). Spear Phishing: In Deconstruction Machines (pp. 149–190). https://doi.org/10.5749/j.ctt20vxpw5.9
- Fishman, B., Best, S., Foster, J., Marx, R., Fishman, B., Best, S., ... Marx, R. (2000). Understanding the Various Types of Denial of Service Attack. (February), 1–16.
- Fook Chong, S., & Choo, R. (2011). Introduction to bootstrap. Proceedings ofSingaporeHealthcare,20(3),236–240.https://doi.org/10.1177/201010581102000314
- Fumera, G., Pillai, I., & Roli, F. (2006). Spam filtering based on the analysis of text information embedded into images. *Journal of Machine Learning Research*, 7, 2699–2720.
- Garcia, F. D., Hoepman, J.-H., & Nieuwenhuizen, J. (2004). Spam Filter Analysis. In Security and Protection in Information Processing Systems (Vol. 147, pp. 395– 410). https://doi.org/10.1007/1-4020-8143-X_26
- goleman, daniel; boyatzis, Richard; Mckee, A. (2019). Summary for Policymakers. In Intergovernmental Panel on Climate Change (Ed.), *Climate Change 2013 - The Physical Science Basis* (Vol. 53, pp. 1–30). https://doi.org/10.1017/CBO9781107415324.004
- Gómez, D. O., Casas, F., Inzunza, J. A., & Costa, P. A. (2017). School and Neighborhood: Influences of Subjective Well-Being in Chilean Children. https://doi.org/10.1007/978-3-319-55601-7_8
- Google. (2019). YouTube Community Guidelines enforcement. *Google Transparency Report*, (August), 2017–2018. Retrieved from https://transparencyreport.google.com/youtube-policy/removals
- Gordon, S. (1995). Internet 101. Computers & Security, 14(7), 599–604. https://doi.org/10.1016/0167-4048(96)81666-X

- Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17, 26–32. https://doi.org/10.1016/j.procs.2013.05.005
- Hans, K., Ahuja, L., & K. Muttoo, S. (2014). Approaches for Web Spam Detection. International Journal of Computer Applications, 101(1), 38–44. https://doi.org/10.5120/17655-8467
- Harvey, K. (2014). Social Media, Definition and Classes of. In *Encyclopedia of Social Media and Politics*. https://doi.org/10.4135/9781452244723.n485
- Hayati, P., Potdar, V., Talevski, A., Firoozeh, N., Sarenche, S., & Yeganeh, E. A. (2010). Definition of spam 2.0: New spamming boom. 4th IEEE International Conference on Digital Ecosystems and Technologies, (May), 580–584. https://doi.org/10.1109/DEST.2010.5610590
- Heavner, S. B., Hardy, S. M., White, D. R., McQueen, C. T., Prazma, J., & Pillsbury,
 H. C. (2001). Image Feature Extraction Techniques and Their Applications for
 CBIR and Biometrics Systems. *Otolaryngology–Head and Neck Surgery*, 125(3),
 123–129. https://doi.org/10.1067/mhn.2001.116448
- Hidalgo, J. M. G., Bringas, G. C., Sánz, E. P., & García, F. C. (2006). Content based SMS spam filtering. *Proceedings of the 2006 ACM Symposium on Document Engineering*, *DocEng* 2006, 2006(June 2014), 107–114. https://doi.org/10.1145/1166160.1166191
- Houvardas, J., & Stamatatos, E. (2006). N-Gram Feature Selection for Authorship Identification. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 4183 LNCS (pp. 77–86). https://doi.org/10.1007/11861461_10
- Hussain, N., Turab Mirza, H., Rasool, G., Hussain, I., & Kaleem, M. (2019). Spam Review Detection Techniques: A Systematic Literature Review. *Applied Sciences*, 9(5), 987. https://doi.org/10.3390/app9050987
- Iii, S. (2009). K-Fold Cross-Validation.
- Issac, B., Chiong, R., & Jacob, S. M. (2006). Analysis of phishing attacks and countermeasures. Managing Information in the Digital Economy: Issues and Solutions - Proceedings of the 6th International Business Information Management Association Conference, IBIMA 2006, 339–346.
- Jackman, W. M., & Roberts, P. (2014). Students' Perspectives on YouTube Video Usage as an E-Resource in the University Classroom. *Journal of Educational*

Technology Systems, 42(3), 273-296. https://doi.org/10.2190/et.42.3.f

- Jothi, G., & H, H. I. (2012). Soft Set Based Feature Selection Approach for Lung Cancer Images. *Ijser*, 3(10), 7.
- Kalaskar, K., & Dhore, M. (2012). Preprocessing Challenges in Document Image Analysis. IJCA Proceedings on National Conference on Recent Trends in Computing, NCRTC(9), 25–30.
- Kanaris, I., Kanaris, K., Houvardas, I., & Stamatatos, E. (2006). Words vs. character *n*-grams for anti-spam filtering. XX(X), 1–20.
- Kant, R., Sengamedu, S. H., & Kumar, K. S. (2012). Comment spam detection by sequence mining. Proceedings of the Fifth ACM International Conference on Web Search and Data Mining - WSDM '12, 183. https://doi.org/10.1145/2124295.2124318
- Kaspersky lab. (2018). *Kaspersky Security Bulletin 2018 STATISTICS*. 25. Retrieved from https://go.kaspersky.com/rs/802-IJN-240/images/KSB_statistics_2018_eng_final.pdf
- Khan, S., Mast, N., Loo, K.-K., & Salahuddin, A. (2008). Passive Security Threats and Consequences in IEEE 802.11 Wireless Mesh Networks. *Researchgate*, 2(3), 4–
 8. Retrieved from http://dblp.unitrier.de/db/journals/jdcta/jdcta2.html#KhanMLS08
- Khawandi, S., Abdallah, F., & Ismail, A. (2019). A Survey On Image Spam Detection Techniques. *Computer Science & Information Technology (CS & IT)*, (January), 13–27. https://doi.org/10.5121/csit.2019.90102
- Krishnamurthy, V. (2015). Internet spam threats and email exploitation A scuffle with inbox attack. (January 2014). https://doi.org/10.6088/ijaser.030400015
- Kumar, G., & Bhatia, P. K. (2014). A Detailed Review of Feature Extraction in Image Processing Systems. 2014 Fourth International Conference on Advanced Computing & Communication Technologies, (February), 5–12. https://doi.org/10.1109/ACCT.2014.74
- Kumar, K. (2016). Intrusion Detection Using Soft Computing Techniques. 6(July), 153–169.
- Kumar, S. V. K., & Kiruthika, P. (2015). An Overview of Classification Algorithm in Data mining. *International Journal of Advanced Research in Computer and Communication* Engineering, 4(12), 255–257. https://doi.org/10.17148/IJARCCE.2015.41259

- Kumar, V. (2014). Feature Selection: A literature Review. *The Smart Computing Review*, 4(3). https://doi.org/10.6029/smartcr.2014.03.007
- Ladha, L., & Deepa, T. (2011). Feature Selection Methods And Algorithms. *International Journal on Computer Science and Engineering*, 3(5), 1787–1797. Retrieved from http://journals.indexcopernicus.com/abstract.php?icid=945099
- Larkin, E. (2005). Spear phishing. PC World (San Francisco, CA), 23(11), 97. https://doi.org/10.5749/j.ctt20vxpw5.9
- Lassification, C. (2011). M Achine L Earning M Ethods for S Pam E- Mail. *Journal* of Computer Science, 3(1), 173–184.
- Lee, I. H., Lushington, G. H., & Visvanathan, M. (2011). A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *Journal* of Clinical Bioinformatics, 1(1), 11. https://doi.org/10.1186/2043-9113-1-11
- Li, Y., Pan, Q., Wang, S., Yang, T., & Cambria, E. (2018). A Generative Model for category text generation. *Information Sciences*, 450(March), 301–315. https://doi.org/10.1016/j.ins.2018.03.050
- Lorentzon, M., & Wallenberg, M. (2017). Feature extraction for image selection using machine learning Feature extraction for image selection using machine learning.
- Ma, Q. (2013). The process and characteristics of phishing attacks A small international trading company case study. *Journal of Technology Research*, 4, 1–16. Retrieved from http://search.proquest.com.library.capella.edu/docview/1460848773?accountid=
 - 27965
- Michelle. (2019). What Is Spam? 2008–2010. Retrieved from https://www.cisco.com/c/en/us/products/security/email-security/what-isspam.html
- Mob, E. I. (1999). Computer viruses demystified. (October), 72.
- MyCert. (2017). mycert-incident-statistics (2).pdf.
- MyCert. (2018). mycert-incident-statistics (1).pdf.
- MyCert. (2019). mycert-incident-statistics.pdf.
- Nath, A., Agarwal, S., & Ghosh, A. (2016). Classification of Machine Learning Algorithms. *International Journal of Innovatice Research in Advanced Engineering*, 3(March), 6–11.
- Nayak, J., Naik, B., & Behera, H. S. (2015). A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges. *International*

Journal of Database Theory and Application, 8(1), 169–186. https://doi.org/10.14257/ijdta.2015.8.1.18

Ogada, K., Mwangi, W., & Wilson, C. (2015). N-gram Based Text Categorization Method for Improved Data Mining. *Journal of Information Engineering and Applications*, 5(8), 35–44.

Orhan, F. (2018). Phishing Got Darker. and Smarter.

Padmavathi, G., & Divya, S. (2013). A Survey on Various Security Threats and Classification of Malware Attacks, Vulnerabilities and Detection Techniques. *The International Journal of Computer Science & Applications (TIJCSA)*, 2(04), 66–72. Retrieved from https://www.semanticscholar.org/paper/A-Survey-on-Various-Security-Threats-and-of-Malware-Padmavathi-

Divya/4be826fe79be9986dc6e7a8e0e0cac491a5b9540#extracted

- Pathical, S., & Serpen, G. (2010). Comparison of subsampling techniques for random subspace ensembles. 2010 International Conference on Machine Learning and Cybernetics, ICMLC 2010, I(August 2010), 380–385. https://doi.org/10.1109/ICMLC.2010.5581032
- Pawar, M. V., & Anuradha, J. (2015). Network security and types of attacks in
network.ProcediaComputerScience,48(C),503–506.https://doi.org/10.1016/j.procs.2015.04.126
- Perumal, S., & Velmurugan, T. (2018). Preprocessing by contrast enhancement techniques for medical images. *Int J Pure Appl Math*, 118(18), 3681–3688.

Pradhan, A. (2017). SUPPORT VECTOR MACHINE-A Survey. (April 2015).

- Qiang Zhang, Chenwei Liu, Shangru Zhong, & Kai Lei. (2017). Spam comments detection with self-extensible dictionary and text-based features. 2017 IEEE Symposium on Computers and Communications (ISCC), 1225–1230. https://doi.org/10.1109/ISCC.2017.8024692
- Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *Eurasip Journal on Advances in Signal Processing*, 2016(1). https://doi.org/10.1186/s13634-016-0355-x
- Renuka, D. K., Hamsapriya, T., Chakkaravarthi, M. R., & Surya, P. L. (2011). Spam Classification Based on Supervised Learning Using Machine Learning Techniques. 2011 International Conference on Process Automation, Control and Computing, 6948(December), 1–7. https://doi.org/10.1109/PACC.2011.5979035

Rosi, S. (2018). SPAM AND EMAIL DETECTION IN BIG DATA PLATFORM

USING. 7(4), 53–58.

- S.ANITHA, D. V. R. (2010). Comparison of Image Preprocessing Techniques for Textile Texture Images. *International Journal of Engineering Science and Technology*, 2(12), 7619–7625.
- S.Mangrulkar, N., R. Bhagat Patil, A., & S. Pande, A. (2014). Network Attacks and Their Detection Mechanisms: A Review. *International Journal of Computer Applications*, 90(9), 37–39. https://doi.org/10.5120/15606-3154
- Samsudin, N. M., Mohd Foozy, C. F. B., Alias, N., Shamala, P., Othman, N. F., & Wan Din, W. I. S. (2019). Youtube spam detection framework using naïve bayes and logistic regression. *Indonesian Journal of Electrical Engineering and Computer Science*, 14(3), 1508. https://doi.org/10.11591/ijeecs.v14.i3.pp1508-1517
- Saptoro, A., Tadé, M. O., & Vuthaluru, H. (2012). A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models. *Chemical Product and Process Modeling*, 7(1). https://doi.org/10.1515/1934-2659.1645
- Sathiaseelan, M. U. M. D. J. G. R. (2017). Text Mining: Survey on Techniques and Applications. *International Journal of Science and Research (IJSR)*, 6(6), 1660– 1664. Retrieved from https://www.ijsr.net/archive/v6i6/ART20174656.pdf
- Science, C., Handayani, A. S., Nurmaini, S., Yani, I., & Husni, N. L. (2017). Analysis for Swarm Robot Coordination using Fuzzy Logic. 5(3), 401–408. https://doi.org/10.11591/ijeecs.v5.i3.pp401-408
- Setia, L., & Burkhardt, H. (2006). Feature Selection for Automatic Image Annotation. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 4174 LNCS (pp. 294–303). https://doi.org/10.1007/11861898_30
- Shahzad, F., Pasha, M., & Ahmad, A. (2017). A Survey of Active Attacks on Wireless Sensor Networks and their Countermeasures. 14(12), 54–65. Retrieved from http://arxiv.org/abs/1702.07136
- Shojaee, S., Murad, M. A. A., Azman, A. Bin, Sharef, N. M., & Nadali, S. (2013). Detecting deceptive reviews using lexical and syntactic features. 2013 13th International Conference on Intellient Systems Design and Applications, 53–58. https://doi.org/10.1109/ISDA.2013.6920707

Silpa, K. S., & Irshad, M. (2018). A survey of lexical simplification. Emerging Trends

in Engineering, Science and Technology for Society, Energy and Environment -Proceedings of the International Conference in Emerging Trends in Engineering, Science and Technology, ICETEST 2018, 60, 785–791. https://doi.org/10.1613/jair.5526

- Singh, P. R., Pandita, R., Kalyanaraman, K., & Chhabra, G. S. (2018). A survey on spam detection techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(12), 171–175. https://doi.org/10.17148/IJARCCE
- Skuratov, V., Kuzmin, K., Nelin, I., & Sedankin, M. (2019). Application of Kohonen neural networks to search for regions of interest in the detection and recognition of objects. *Eastern-European Journal of Enterprise Technologies*, 3(9–99), 41– 48. https://doi.org/10.15587/1729-4061.2019.166887
- Sosa, J. N. (2010). Spam Classification Using Machine Learning Techniques -Sinespam. (August), 427.
- Subramaniam, T., Jalab, H. A., & Taqa, A. Y. (2010). Overview of textual anti-spam filtering techniques. *International Journal of Physical Sciences*, 5(12), 1869– 1882.
- Tang, J. (2012). Generative and Discriminative Models ML as Searching Hypotheses Space.
- Tian, D. P. (2013). A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, 8(4), 385–395.
- Tripathi, N., Swarnkar, M., & Hubballi, N. (2018). DNS spoofing in local networks made easy. 11th IEEE International Conference on Advanced Networks and Telecommunications Systems, ANTS 2017, (October 2017), 1–6. https://doi.org/10.1109/ANTS.2017.8384122
- Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57(March), 117–126. https://doi.org/10.1016/j.eswa.2016.03.028
- Uysal, A. K. (2018). Feature Selection for Comment Spam Filtering on YouTube.
- Venkatesh, B., & Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, 19(1), 3–26. https://doi.org/10.2478/cait-2019-0001
- Wang, D., Irani, D., & Pu, C. (2011). A social-spam detection framework. Proceedings

of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference on - CEAS '11, 46–54. https://doi.org/10.1145/2030376.2030382

- Wang, S., Tang, J., & Liu, H. (2016). Encyclopedia of Machine Learning and Data Mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* and Data Mining. https://doi.org/10.1007/978-1-4899-7502-7
- Wang, Z., Josephson, W., Lv, Q., Charikar, M., & Li, K. (2007). Filtering image spam with near-duplicate detection. 4th Conference on Email and Anti-Spam, CEAS 2007.
- Wei, Z., Miao, D., Chauchat, J.-H., Zhao, R., & Li, W. (2009). N-grams based feature selection and text representation for Chinese Text Classification. *International Journal of Computational Intelligence Systems*, 2(4), 365–374. https://doi.org/10.1080/18756891.2009.9727668
- Wolf, M., Sims, J., & Yang, H. (2017). Social Media? What Social Media? Sage Jornal, 17.
- Wu, J., & Deng, T. (2008). Research in Anti-Spam Method Based on Bayesian Filtering. 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, 2, 887–891. https://doi.org/10.1109/PACIIA.2008.180
- Wuensch, K. L. (2011). Chi-Square Tests. In International Encyclopedia of Statistical Science (pp. 252–253). https://doi.org/10.1007/978-3-642-04898-2_173
- Yan Gao, Ming Yang, Xiaonan Zhao, Bryan Pardo, Ying Wu, Pappas, T. N., & Alok Choudhary. (2008). Image spam hunter. 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 1765–1768. https://doi.org/10.1109/ICASSP.2008.4517972
- Ye, Y., Li, T., Adjeroh, D., & Iyengar, S. S. (2017). A Survey on Malware Detection Using Data Mining Techniques. ACM Computing Surveys, 50(3), 1–40. https://doi.org/10.1145/3073559
- Yusof, Y., & Sadoon, O. H. (2017). Detecting Video Spammers in Youtube Social Media. ICOCI Kuala Lumpur. Universiti Utara Malaysia, (082), 25–27. Retrieved from http://www.uum.edu.my
- ZDH. (2018). A Brief Survey of Text Simplification. (October), 1–5.
- Zhou, X., & Wang, J. (2015). Feature Selection for Image Classification Based on a New Ranking Criterion. *Journal of Computer and Communications*, 03(03), 74– 79. https://doi.org/10.4236/jcc.2015.33013

APPENDIX

A. Gantt Chart part 1

D	0	Task Mode	Task Name	Duration	Start	Finish	Predecessors
1		N.	YOUTUBE SPAM CLASSIFICATION USING WORD FREQUENCIES	70 days	Mon 9/9/19	Fri 12/13/19	
2	-	-	Planning	25 days	Mon 9/9/19	Fri 10/11/19	}
3			Identify title, problem statement and scope,	5 days	Mon 9/9/19	Fri 9/13/19	
4	(MIL)	an He	Study and research the literature review. Write project proposal to supervisor.	5 days	Mon 9/16/19	Fri 9/20/19	3
5	E E	-	Proposal accept	5 days	Mon 9/23/19	Fri 9/27/19	4
6	180	A SAIN	Identify title, problem statement, objective and scope of	5 days	Mon 9/30/19	Fri 10/4/19	5
7		Plo	Chapter 1 is done and submit to	5 days	Mon 10/7/19	En 10/11/19	اوىيۇ
		NIVE	RS supervisor for evaluation.	IKAL	MALAYS	SIA MEI	LAKA

B. Gantt Chart part 2

8	-	Analysis	20 days	Mon 10/14	/1Fri 11/8/19	2
9	R.	Studies on related work and previous research and finding taxonomy of spam classification	5 days	Mon 10/14/19	Fri 10/18/19	7
10	-	Study methodology on previous research.	5 days	Mon 10/21/19	Fri 10/25/19	9
11	-	MID SEMESTER BREAK	5 days	Mon 10/28/19	Fri 11/1/19	10
12	WAL MAI	Information Av collection an analysis.	5 days d	Mon 11/4/19	Fri 11/8/19	11
13	N	Design 🖇	25 days	Mon 11/11	/1Fri 12/13/19	8
14	L LIGUNATION	Design the project and choose the tools for	5 days	Mon 11/11/19	Fri 11/15/19	12
15	املاك	Design the environment for implementation	5 days	Mon 11/18/19	Fri 11/22/19	او ب
16	UMPVER	RS Write and finalize proje report.	ct	11/25/19	/Fri 11/29/19	1 ₹.A
17	-	Submit proje report to supervisor.	ect 5 days	Mon 12/2/19	Fri 12/6/19	16
18	-	Schedule the	Pr 5 days	Mon 12/9/1	9Fri 12/13/19	17
Project Date: T	t gancaattt me Thu 12/12/19	gat Task Split Milestor	1e	•	Summary Project Summ Inactive Task	ary I

C. Gantt Chart part 3



Page 1

89