

# **YOUTUBE SPAM DETECTION USING ENSEMBLE METHOD**



**UNIVERSITI TEKNIKAL MALAYSIA MELAKA**

## BORANG PENGESAHAN STATUS LAPORAN

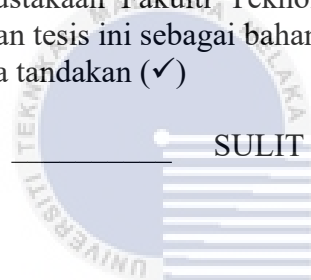
JUDUL: [YOUTUBE SPAM DETECTION USING ENSEMBLE METHOD]

SESI PENGAJIAN: [2020 / 2021]

Saya: \_\_\_\_\_ SYAZA LIYANA BINTI MUHAMAD SHAPEE \_\_\_\_\_  
(HURUF BESAR)

mengaku membenarkan tesis Projek Sarjana Muda ini disimpan di Perpustakaan Universiti Teknikal Malaysia Melaka dengan syarat-syarat kegunaan seperti berikut:

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



\_\_\_\_\_ SULIT

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

\_\_\_\_\_ TERHAD

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

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

\_\_\_\_\_ ✓ \_\_\_\_\_ TIDAK TERHAD

(TANDATANGAN PELAJAR)

Alamat tetap: NO.3, SEKUNTUM 21,  
TAMAN BUKIT DAHLIA 81700  
PASIR GUDANG, JOHOR

Tarikh: 7 SEPTEMBER 2021

(TANDATANGAN PENYELIA)

MR. NOR AZMAN BIN MAT ARIFF

Nama Penyelia

Tarikh: 8/9/2021

## **YOUTUBE SPAM DETECTION USING ENSEMBLE METHOD**

SYAZA LIYANA BINTI MUHAMAD SHAPEE



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

اويؤر ستي بيكنيكل مليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY  
UNIVERSITI TEKNIKAL MALAYSIA MELAKA  
2021**

## DECLARATION

I hereby declare that this project report entitled  
**YOUTUBE SPAM DETECTION USING ENSEMBLE METHOD**  
is written by me and is my own effort and that no part has been plagiarized  
without citations.

STUDENT:

  
SYAZA LIYANA BINTI MUHAMAD SHAPEE

Date : 7 SEPTEMBER 2021



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

SUPERVISOR :

  
MR. NOR AZMAN BIN MAT ARIFF

Date :

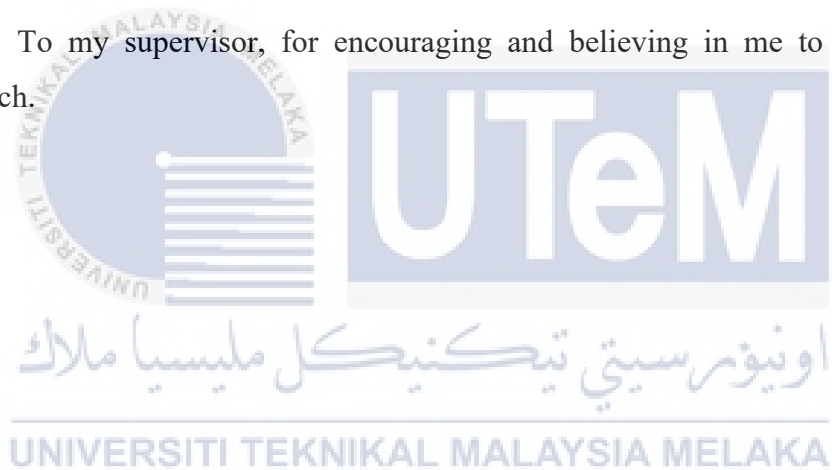
8/9/2021

## DEDICATION

To my beloved parents, Zaleha binti Ali and Muhamad Shapee bin Ghazali who inspired me to be strong despite of many obstacles in life, for their prayers and their overwhelming support morally and financially. My sisters and brother, Syaza Hazirah, Syaza Najihah and Muhammad Adam have never left my side and are very special.

To my fellow friends, for being there for me throughout the entire bachelor program and their cooperation while conducting the research.

To my supervisor, for encouraging and believing in me to complete this research.



## ACKNOWLEDGEMENTS

All praises be to Almighty Allah S.W.T who has blessed me with the belief, strength and capabilities to understand, learn and complete this research. Peace and prayers be upon our most beloved Prophet Muhammad S.A.W, the most beautiful soul, whose sayings, actions and stories have deeply inspired me enough to believe that there are no limitations to what I can achieve when we are fully committed to accomplish something, knowing that Allah is on my side.

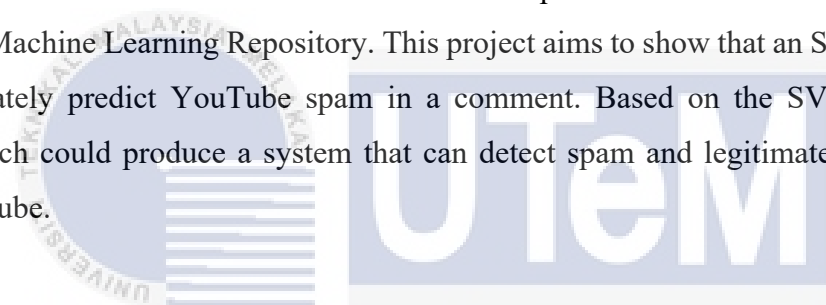
I also admire the help and the guidance of my supervisor, Mr. Nor Azman bin Mat Ariff for his guidance, encouragement and patience from all aspects during the preparation of this research are highly appreciated.

I am blessed to have had such wonderful, loving and supporting parents, Zaleha binti Ali and Muhammad Shapee bin Ghazali for the education they gave me at home as I was growing up and the education, they paid for till I graduated. They have been my pillar of strength and till this very day, every small achievement I make, they always want to be the first to know and to congratulate.

To my fellow friends who have helped in the strenuous process in collecting information and preparing this research. May Allah bless you all for your patience and selfless commitment.

## ABSTRACT

The number of YouTube users is constantly rising. However, such success is not without its drawbacks. Spam has become a common form of attack and threat, and most YouTube users are unaware of it. Receiving and being overwhelmed with unnecessary spam regularly has become one of the most internet-disruptive topics in today's world. The Support Vector Machine (SVM) is used in this study to develop a YouTube detection framework. The YouTube spam datasets were obtained from the UCI Machine Learning Repository. This project aims to show that an SVM model can accurately predict YouTube spam in a comment. Based on the SVM model, this research could produce a system that can detect spam and legitimate comments on YouTube.



اونيورسيتي تيكنيكل مليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

## ABSTRAK

Bilangan pengguna YouTube terus meningkat. Namun, kejayaan itu bukan tanpa kekurangannya. Spam telah menjadi bentuk serangan dan ancaman yang biasa, dan kebanyakan pengguna YouTube tidak menyedarinya. Menerima dan dibanjiri dengan spam yang tidak perlu secara berkala telah menjadi salah satu topik yang mengganggu internet di dunia sekarang. Mesin Vektor Sokongan (SVM) digunakan dalam kajian ini untuk mengembangkan kerangka pengesanan YouTube. Set data spam YouTube diperoleh dari UCI Machine Learning Repository. Projek ini bertujuan untuk menunjukkan bahawa model SVM dapat meramalkan spam YouTube dengan tepat dalam komen. Berdasarkan model SVM, penyelidikan ini dapat menghasilkan sistem yang dapat mengesan spam dan komen yang sah di YouTube.

اونيورسيتي تيكنيكل مليسيا ملاك

UNIVERSITI TEKNIKAL MALAYSIA MELAKA



## TABLE OF CONTENTS

	<b>PAGE</b>
<b>DECLARATION</b>	<b>II</b>
<b>DEDICATION</b>	<b>III</b>
<b>ACKNOWLEDGEMENTS</b>	<b>IV</b>
<b>ABSTRACT</b>	<b>V</b>
<b>ABSTRAK</b>	<b>VI</b>
<b>TABLE OF CONTENTS</b>	<b>VII</b>
<b>LIST OF TABLES</b>	<b>XVI</b>
<b>LIST OF FIGURES</b>	<b>1</b>
<b>LIST OF ABBREVIATIONS</b>	<b>5</b>
<b>LIST OF ATTACHMENTS</b>	<b>6</b>
<b>CHAPTER 1: INTRODUCTION</b>	<b>7</b>
1.1 Introduction	7
1.2 Problem Statement (PS)	8
1.3 Project Questions (PQ)	9
1.4 Project Objective (PO)	9
1.5 Project Scope	10
1.6 Project Contribution	10
1.7 Report Organization	11
1.7.1 Chapter I: Introduction	11

1.7.2	Chapter II: Literature Review	11
1.7.3	Chapter III: Project Methodology	11
1.7.4	Chapter IV: Analysis and Design	11
1.7.5	Chapter V: Implementation	11
1.7.6	Chapter VI: Discussion	11
1.7.7	Chapter VII: Project Conclusion	12
1.8	Conclusion	12
<b>CHAPTER 2: LITERATURE REVIEW</b>		<b>13</b>
2.1	Introduction	13
2.2	General Categories of Internet Security Attack (ISA)	15
2.2.1	ISA Definition	15
2.2.2	Denial of Service Attack (DoS)	15
2.2.3	Phishing Attack	15
2.2.3.1	Clone Phishing	16
2.2.3.2	Spear Phishing	16
2.2.3.3	DNS Base Phishing	16
2.2.4	Spam Attack	17
2.2.5	Malware	17
2.2.6	Virus	17
2.3	Classification of ISA	18
2.3.1	Active Attack	18
2.3.2	Passive Attack	18
2.4	Spam	19
2.4.1	Spam Definition	19
2.4.2	Spam Type	19

2.4.2.1	E-mail	19
2.4.2.2	Web Spam	20
2.4.2.3	Short Message Service (SMS) Spam	21
2.4.2.4	Image Spam	21
2.4.2.5	YouTube Spam	22
2.4.3	Spam Detection Technique	23
2.4.4	Spam Analysis Technique	24
2.5	Machine Learning	25
2.5.1	Machine Learning Definition	25
2.5.2	Dataset	26
2.5.3	Data Preprocessing	26
2.5.3.1	Definition	26
2.5.3.2	Preprocessing Type	26
2.5.4	Feature Extraction	27
2.5.4.1	N-Gram	28
2.5.4.2	Lexical Features	28
2.5.4.3	Ensemble Method	29
2.5.5	Data Splitting and Validation	32
2.5.5.1	Random Subsampling	32
2.5.5.2	Cross-validation	33
2.5.5.3	Bootstrapping	33
2.5.6	Feature Selection	33
2.5.6.1	Feature Selection Definitions	33

2.5.6.2	Feature Selection Type	34
2.5.7	Classification	36
2.5.8	Classification Type	36
2.5.8.1	Generative	36
2.5.8.2	Discriminative	38
2.6	Critical Review	39
2.6.1	Previous Research on Spam	39
2.6.1.1	Research Paper I	39
2.6.1.2	Research Paper II	39
2.6.1.3	Research Paper III	40
2.6.2	Previous Research on YouTube Spam	41
2.6.2.1	Research Paper I	41
2.6.2.2	Research Paper II	41
2.6.2.3	Research Paper III	42
2.6.2.4	Research Paper IV	42
2.7	Conclusion	44
<b>CHAPTER 3: PROJECT METHODOLOGY</b>		<b>45</b>
3.1	Introduction	45
3.2	Methodology	45
3.2.1	Previous Research	46
3.2.2	Information Gathering	47
3.2.3	Define Scope	47
3.2.4	Design and Implementation	47

3.2.5	Testing and Evaluation of Model	47
3.2.6	Documentation	47
3.3	Project Schedule and Milestones	48
3.3.1	Project Flowchart	48
3.3.2	Project Milestones	49
3.3.3	Project Gantt Chart	50
3.4	Requirement Analysis	50
3.4.1	Software Requirement	50
3.4.2	Hardware Requirement	51
3.5	Conclusion	52
<b>CHAPTER 4: ANALYSIS AND DESIGN</b>		<b>53</b>
4.1	Introduction	53
4.2	Problem Analysis	53
4.3	Project Design	53
4.3.1	Dataset	54
4.3.2	Data Preprocessing	55
4.3.3	Feature Extraction	58
4.3.4	Generate Bag of Word (BOW) Feature Vector	59
4.3.5	Data Splitting and Validation	59
4.3.6	Feature Selection	60
4.3.7	Normalization	62
4.3.8	Classification	62
4.3.9	Post-Classification	66
4.4	Conclusion	66
<b>CHAPTER 5: IMPLEMENTATION</b>		<b>67</b>

5.1	Introduction	67
5.2	Software Development Environment Setup	67
5.3	Process Module	68
5.3.1	Collection of Dataset	68
5.3.2	Data Preprocessing	68
5.3.2.1	Eminem Dataset Preprocessing	68
5.3.2.2	Psy Dataset Preprocessing	72
5.3.2.3	Shakira Dataset Preprocessing	76
5.3.2.4	LMFAO Dataset Preprocessing	79
5.3.2.5	Katy Perry Dataset Preprocessing	83
5.3.3	Feature Extraction	86
5.3.3.1	Eminem Dataset Feature Extraction	87
5.3.3.2	Psy Dataset Feature Extraction	89
5.3.3.3	Shakira Dataset Feature Extraction	90
5.3.3.4	LMFAO Dataset Feature Extraction	92
5.3.3.5	Katy Perry Dataset Feature Extraction	94
5.3.4	Feature Vector	95
5.3.4.1	Dataset Eminem Feature Vector	96
5.3.4.2	Dataset Psy Feature Vector	99
5.3.4.3	Dataset Shakira Feature Vector	102
5.3.4.4	Dataset LMFAO Feature Vector	106
5.3.4.5	Dataset Katy Perry Feature Vector	109

5.3.5	Data Traing and Testing	113
5.3.5.1	Splitting Data	113
5.3.5.2	Model File	115
5.3.5.3	Predict File	116
5.4	Script Execution	116
5.4.1	Eminem Dataset Scripts	116
5.4.2	Psy Dataset Scripts	118
5.4.3	Shakira Dataset Scripts	119
5.4.4	LMFAO Dataset Scripts	120
5.4.5	Katy Perry Dataset Scripts	121
5.5	Result	122
5.5.1	Eminem Dataset Result	123
5.5.1.1	Attribute	123
5.5.1.2	10 Runs	123
5.5.1.3	Accuracy Table	125
5.5.2	Psy Dataset Result	126
5.5.2.1	Attribute	126
5.5.2.2	10 Runs	127
5.5.2.3	Accuracy Table	128
5.5.3	Shakira Dataset Result	130
5.5.3.1	Attribute	130
5.5.3.2	10 Runs	131
5.5.3.3	Accuracy Table	132

5.5.4	LMFAO Dataset Result	133
5.5.4.1	Attribute	133
5.5.4.2	10 Runs	134
5.5.4.3	Accuracy Table	135
5.5.5	Katy Perry Dataset Result	137
5.5.5.1	Attribute	137
5.5.5.2	10 Runs	138
5.5.5.3	Accuracy Table	139
5.5.6	Summary Dataset Result	140
5.6	Conclusion	141
<b>CHAPTER 6: DISCUSSION</b>		<b>142</b>
6.1	Introduction	142
6.2	Discussion of Project	142
6.3	Discussion on The Newly Proposed Method	143
6.4	Conclusion	144
<b>CHAPTER 7: PROJECT CONCLUSION</b>		<b>145</b>
7.1	Introduction	145
7.2	Project Summary	145
7.3	Project Constraint	146
7.4	Project Contribution	146
7.5	Project Limitation	146
7.6	Future Work	146



7.7	Conclusion	146
	<b>REFERENCES</b>	<b>147</b>
	<b>APPENDIX</b>	<b>154</b>



## LIST OF TABLES

<b>Table 1.1: Problem Statement</b>	<b>8</b>
<b>Table 1.2 Summary of Project Question</b>	<b>9</b>
<b>Table 1.3 Summary of Project Objective</b>	<b>9</b>
<b>Table 1.4: Project Contribution</b>	<b>10</b>
<b>Table 2.1: SPAM Literature</b>	<b>40</b>
<b>Table 2.2: YouTube SPAM literature</b>	<b>42</b>
<b>Table 3.1 Project Milestone</b>	<b>49</b>
<b>Table 3.2: Software Requirement for the Project</b>	<b>51</b>
<b>Table 3.3: Hardware Requirement of the Project</b>	<b>51</b>
<b>Table 4.1 Description of Dataset</b>	<b>55</b>
<b>Table 4.2: Advantage and Disadvantage of Rules</b>	<b>58</b>
<b>Table 4.3: Type of SVM Kernel</b>	<b>65</b>
<b>Table 5.1: Generate Ngram Eminem</b>	<b>87</b>
<b>Table 5.2: Feature Descriptor Eminem</b>	<b>88</b>
<b>Table 5.3: Parameter Details of Generate Ngram Psy</b>	<b>89</b>
<b>Table 5.4: Feature Descriptor Psy</b>	<b>90</b>
<b>Table 5.5: Generate Ngram Shakira</b>	<b>91</b>
<b>Table 5.6: Feature Descriptor Shakira</b>	<b>92</b>
<b>Table 5.7: Parameter Details of Generate Ngram LMFAO</b>	<b>92</b>
<b>Table 5.8: Feature Descriptor LMFAO</b>	<b>93</b>
<b>Table 5.9: Generate Ngram Katy Perry</b>	<b>94</b>
<b>Table 5.10: Feature Descriptor Katy Perry</b>	<b>95</b>
<b>Table 5.11: Details Parameter of Feature Vector Eminem</b>	<b>96</b>
<b>Table 5.12: Generate Feature Vector Eminem for Both Experiments</b>	<b>97</b>
<b>Table 5.13: Details Parameter of Convert CSV to ARFF Eminem</b>	<b>98</b>
<b>Table 5.14: Details Parameter of Feature Vector Psy</b>	<b>99</b>

<b>Table 5.15: Generate Feature Vector Psy for Both Experiments</b>	<b>100</b>
<b>Table 5.16: Details Parameter of Convert CSV to ARFF Psy</b>	<b>102</b>
<b>Table 5.17: Details Parameter of Feature Vector Shakira</b>	<b>103</b>
<b>Table 5.18: Generate Feature Vector Shakira for Both Experiments</b>	<b>104</b>
<b>Table 5.19: Details Parameter of Convert CSV to ARFF Shakira</b>	<b>105</b>
<b>Table 5.20: Details Parameter of Feature Vector LMFAO</b>	<b>106</b>
<b>Table 5.21: Generate Feature Vector LMFAO for Both Experiments</b>	<b>107</b>
<b>Table 5.22: Details Parameter of Convert CSV to ARFF LMFAO</b>	<b>109</b>
<b>Table 5.23: Details Parameter of Feature Vector Katy Perry</b>	<b>110</b>
<b>Table 5.24: Generate Feature Vector Katy Perry for Both Experiments</b>	<b>111</b>
<b>Table 5.25: Details Parameter of Convert CSV to ARFF Katy Perry</b>	<b>112</b>
<b>Table 5.26: Sorted Number of Instances of Each Dataset</b>	<b>113</b>
<b>Table 5.27: Training and Testing Size</b>	<b>114</b>
<b>Table 5.28: Parameter description Eminem</b>	<b>117</b>
<b>Table 5.29: Parameter description Psy</b>	<b>118</b>
<b>Table 5.30: Parameter description Shakira for experiment 1</b>	<b>119</b>
<b>Table 5.31: Parameter description LMFAO for experiment 1</b>	<b>121</b>
<b>Table 5.32: Parameter description Katy Perry for experiment 1</b>	<b>122</b>
<b>Table 5.33: 80% of Total Instances Attributes Eminem</b>	<b>123</b>
<b>Table 5.34: Single Classifier of Eminem Accuracy</b>	<b>125</b>
<b>Table 5.35: 80% of Total Instances Attributes Psy</b>	<b>127</b>
<b>Table 5.36: Single Classifier of Psy Accuracy</b>	<b>128</b>
<b>Table 5.37: 80% of Total Instances Attributes Shakira</b>	<b>130</b>
<b>Table 5.38: Single Classifier of Shakira Accuracy</b>	<b>132</b>
<b>Table 5.39: 80% of Total Instances Attributes LMFAO</b>	<b>134</b>
<b>Table 5.40: Single Classifier of LMFAO Accuracy</b>	<b>135</b>
<b>Table 5.41: 80% of Total Instances Attributes Katy Perry</b>	<b>137</b>
<b>Table 5.42: Single Classifier of Katy Perry Accuracy</b>	<b>139</b>
<b>Table 5.43: Dataset Summarization</b>	<b>140</b>
<b>Table 6.1: The Benchmark</b>	<b>143</b>
<b>Table 6.2: Final Result</b>	<b>144</b>

## LIST OF FIGURES

<b>Figure 2.1: Literature Review's Structure</b>	<b>14</b>
<b>Figure 2.2: Type of Spam</b>	<b>19</b>
<b>Figure 2.3: Example E-mail Spam</b>	<b>20</b>
<b>Figure 2.4: Example of SMS Spam</b>	<b>21</b>
<b>Figure 2.5: Example spam comment on YouTube</b>	<b>22</b>
<b>Figure 2.6: Illustration of Bagging Technique in Ensemble Method</b>	<b>30</b>
<b>Figure 2.7: Illustration of Boosting Technique in Ensemble Method</b>	<b>31</b>
<b>Figure 2.8: Wrapper Methods</b>	<b>35</b>
<b>Figure 2.9: Filter Methods</b>	<b>35</b>
<b>Figure 3.1: Framework of the System</b>	<b>46</b>
<b>Figure 3.2: Project Flowchart</b>	<b>48</b>
<b>Figure 4.1: Project Design</b>	<b>54</b>
<b>Figure 4.2: Example of Psy dataset</b>	<b>55</b>
<b>Figure 4.3: Example of features will be use</b>	<b>55</b>
<b>Figure 4.4: Raw Data</b>	<b>56</b>
<b>Figure 4.5: Tokenization</b>	<b>56</b>
<b>Figure 4.6: Token Length</b>	<b>56</b>
<b>Figure 4.7: Case Normalization</b>	<b>57</b>
<b>Figure 4.8: Special Character</b>	<b>57</b>
<b>Figure 4.9: Stop Word Removal</b>	<b>57</b>
<b>Figure 4.10: Stemming</b>	<b>57</b>
<b>Figure 4.11: Subsampling</b>	<b>59</b>
<b>Figure 4.12: SVM Graph</b>	<b>62</b>
<b>Figure 4.13: Hyperplane Placement</b>	<b>63</b>
<b>Figure 4.14: Real Distribution Data Example</b>	<b>64</b>
<b>Figure 4.15: Non-Linear Graph</b>	<b>64</b>
<b>Figure 4.16: SVM with Multidimensional Space</b>	<b>65</b>
<b>Figure 5.1: Process Module</b>	<b>68</b>

<b>Figure 5.2: Data Preprocessing Module</b>	<b>68</b>
<b>Figure 5.3: Sort Dataset Eminem</b>	<b>69</b>
<b>Figure 5.4: Content Dataset Eminem</b>	<b>69</b>
<b>Figure 5.5: Remove Non-ASCII Character</b>	<b>70</b>
<b>Figure 5.6: No Author</b>	<b>70</b>
<b>Figure 5.7: No Date</b>	<b>71</b>
<b>Figure 5.8: Convert To Lowercase</b>	<b>71</b>
<b>Figure 5.9: Sample scriptEminem.bat</b>	<b>72</b>
<b>Figure 5.10: Sample File Created for Eminem</b>	<b>72</b>
<b>Figure 5.11: Sort Dataset Psy</b>	<b>73</b>
<b>Figure 5.12: Content Psy Dataset</b>	<b>73</b>
<b>Figure 5.13: Remove Non-ASCII Character</b>	<b>74</b>
<b>Figure 5.14: No Author</b>	<b>74</b>
<b>Figure 5.15: Convert To Lowercase</b>	<b>75</b>
<b>Figure 5.16: Sample scriptPsy.bat</b>	<b>75</b>
<b>Figure 5.17: Sample File Created for Psy</b>	<b>76</b>
<b>Figure 5.18: Sort Dataset Shakira</b>	<b>76</b>
<b>Figure 5.19: Content Shakira Dataset</b>	<b>77</b>
<b>Figure 5.20: Remove Non-ASCII Character</b>	<b>77</b>
<b>Figure 5.21: No Author</b>	<b>78</b>
<b>Figure 5.22: Convert To Lowercase</b>	<b>78</b>
<b>Figure 5.23: Sample scriptShakira.bat</b>	<b>79</b>
<b>Figure 5.24: Sample File Created for Shakira</b>	<b>79</b>
<b>Figure 5.25: Sort LMFAO Dataset</b>	<b>80</b>
<b>Figure 5.26: Content LMFAO Dataset</b>	<b>80</b>
<b>Figure 5.27: Remove Non-ASCII Character</b>	<b>81</b>
<b>Figure 5.28: No Author</b>	<b>81</b>
<b>Figure 5.29: Convert To Lowercase</b>	<b>82</b>
<b>Figure 5.30: Sample scriptLMFAO.bat</b>	<b>83</b>
<b>Figure 5.31: Sample File Created for LMFAO</b>	<b>83</b>
<b>Figure 5.32: Sort Katy Perry Dataset</b>	<b>84</b>
<b>Figure 5.33: Content Katy Perry Dataset</b>	<b>84</b>
<b>Figure 5.34: Remove Non-ASCII Character</b>	<b>84</b>
<b>Figure 5.35: No Author</b>	<b>85</b>

<b>Figure 5.36: Convert To Lowercase</b>	<b>85</b>
<b>Figure 5.37: Sample scriptKatyPerry.bat</b>	<b>86</b>
<b>Figure 5.38: Sample File Created for Katy Perry</b>	<b>86</b>
<b>Figure 5.39: Generate Ngram Eminem</b>	<b>87</b>
<b>Figure 5.40: Batch File Eminem</b>	<b>88</b>
<b>Figure 5.41: Generate Ngram Psy</b>	<b>89</b>
<b>Figure 5.42: Batch File Psy</b>	<b>90</b>
<b>Figure 5.43: Generate Ngram Shakira</b>	<b>91</b>
<b>Figure 5.44: Batch File Shakira</b>	<b>91</b>
<b>Figure 5.45: Generate Ngram LMFAO</b>	<b>92</b>
<b>Figure 5.46: Batch File LMFAO</b>	<b>93</b>
<b>Figure 5.47: Generate Ngram Katy Perry</b>	<b>94</b>
<b>Figure 5.48: Batch File Katy Perry</b>	<b>95</b>
<b>Figure 5.49: Feature Vector Module</b>	<b>95</b>
<b>Figure 5.50: Batch File for Ngram Feature Vector Eminem</b>	<b>96</b>
<b>Figure 5.51: Batch File Convert CSV to ARFF Eminem</b>	<b>98</b>
<b>Figure 5.52: Batch File for Ngram Feature Vector Psy</b>	<b>99</b>
<b>Figure 5.53: Batch File Convert CSV to ARFF Psy</b>	<b>102</b>
<b>Figure 5.54: Batch File for Ngram Feature Vector Shakira</b>	<b>103</b>
<b>Figure 5.55: Batch File Convert CSV to ARFF Shakira</b>	<b>105</b>
<b>Figure 5.56: Batch File for Ngram Feature Vector LMFAO</b>	<b>106</b>
<b>Figure 5.57: Batch File Convert CSV to ARFF LMFAO</b>	<b>109</b>
<b>Figure 5.58: Batch File for Ngram Feature Vector Katy Perry</b>	<b>110</b>
<b>Figure 5.59: Batch File Convert CSV to ARFF Katy Perry</b>	<b>112</b>
<b>Figure 5.60: Train and Test Module</b>	<b>113</b>
<b>Figure 5.61: tern and Structure Training and Testing Data</b>	<b>115</b>
<b>Figure 5.62: Train Process of Model and Scale</b>	<b>116</b>
<b>Figure 5.63: Process Generating Predict File</b>	<b>116</b>
<b>Figure 5.64: Run code program Eminem experiment 1</b>	<b>117</b>
<b>Figure 5.65: Run code program Eminem experiment 2</b>	<b>117</b>
<b>Figure 5.66: Run code program Psy experiment 1</b>	<b>118</b>
<b>Figure 5.67: Run code program Psy experiment 2</b>	<b>118</b>
<b>Figure 5.68: Run code program Shakira experiment 1</b>	<b>119</b>
<b>Figure 5.69: Run code program Shakira experiment 2</b>	<b>119</b>

<b>Figure 5.70: Run code program LMFAO experiment 1</b>	<b>120</b>
<b>Figure 5.71: Run code program LMFAO experiment 2</b>	<b>120</b>
<b>Figure 5.72: Run code program Katy Perry experiment 1</b>	<b>121</b>
<b>Figure 5.73: Run code program Katy Perry experiment 2</b>	<b>122</b>
<b>Figure 5.74: Example of 10 Runs Experiment for 1 gram by Eminem</b>	<b>124</b>
<b>Figure 5.75: Graph of Experiment 1 Eminem</b>	<b>124</b>
<b>Figure 5.76: Multiple Classifier of Eminem Accuracy</b>	<b>126</b>
<b>Figure 5.77: Example of 10 Runs Experiment for 1 gram by Psy</b>	<b>128</b>
<b>Figure 5.78: Graph of Experiment 1 Psy</b>	<b>128</b>
<b>Figure 5.79: Multiple Classifier of Psy Accuracy</b>	<b>129</b>
<b>Figure 5.80: Example of 10 Runs Experiment for 1 gram by Shakira</b>	<b>131</b>
<b>Figure 5.81: Graph of Experiment 1 Shakira</b>	<b>131</b>
<b>Figure 5.82: Multiple Classifier of Shakira Accuracy</b>	<b>133</b>
<b>Figure 5.83: Example of 10 Runs Experiment for 1 gram by LMFAO</b>	<b>135</b>
<b>Figure 5.84: Graph of Experiment 1 LMFAO</b>	<b>135</b>
<b>Figure 5.85: Multiple Classifier of LMFAO Accuracy</b>	<b>136</b>
<b>Figure 5.86: Example of 10 Runs Experiment for 1 gram by Katy Perry</b>	<b>138</b>
<b>Figure 5.87: Graph of Experiment 1 Katy Perry</b>	<b>139</b>
<b>Figure 5.88: Multiple Classifier of Katy Perry Accuracy</b>	<b>140</b>

## LIST OF ABBREVIATIONS

<b>ASCII</b>	-	<b>American Standard Code For Information Interchange</b>
<b>BOW</b>	-	<b>Bag-Of-Word</b>
<b>CS</b>	-	<b>Chi-Square</b>
<b>DNS</b>	-	<b>Domain Name System</b>
<b>DOS</b>	-	<b>Denial Of Service Attack</b>
<b>Email</b>	-	<b>Electronic Mail</b>
<b>FN</b>	-	<b>False Negatives</b>
<b>FP</b>	-	<b>False Positives</b>
<b>FS</b>	-	<b>Feature Selection</b>
<b>FYP</b>	-	<b>Final Year Project</b>
<b>GIF</b>	-	<b>Graphic Interchange Format</b>
<b>HTML</b>	-	<b>Hypertext Markup Language</b>
<b>IG</b>	-	<b>Information Gain</b>
<b>IDE</b>	-	<b>Integrated Drive Electronics</b>
<b>Weka</b>	-	<b>Waikato Environment for Knowledge Analysis</b>
<b>SVM</b>	-	<b>Support Vector Machine</b>