

**TEXT CLASSIFICATION TECHNIQUES USED TO FACILITATE
CYBER TERRORISM INVESTIGATION**

By

David Allister Simanjuntak



SWISS GERMAN UNIVERSITY
EduTown BSDCity
Tangerang – 15339
Island of Java, Indonesia
www.sgu.ac.id

July 2010

Revised after Thesis Defense on 27 July 2010

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, not material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.

David Allister Simanjuntak

Date

Approved by:

Anto Satriyo Nugroho, Dr.Eng

Date

Charles Lim, Msc., ECSA, ECSP, ECIH, CEH, CEI

Date

Chairman of the Examination Steering Committee

Date

David Allister Simanjuntak

ABSTRACT

TEXT CLASSIFICATION TECHNIQUES USED TO FACILITATE CYBER TERRORISM INVESTIGATION

By

David Allister Simanjuntak

SWISS GERMAN UNIVERSITY

Bumi Serpong Damai

Anto Satriyo Nugroho, Major Advisor

The rising of computer violence, such as Distributed Denial of Service (DDoS), web vandalism, and cyber bullying become more serious issues when they are politically motivated and intentionally conducted to generate fear in society. These kinds of activities are categorized as cyber terrorism. As the number of such cases increase, the availability of information regarding these actions is required to facilitate experts in investigating cyber terrorism. Meanwhile, web mining is one of significant technologies applied to extract information from the Web. In this case, web mining facilitates data acquisition related to cyber terrorism information from the Web. This research aims to create text classification technique based upon number of occurrences of certain relevant words in the term of Cyber Terrorism. This research compared the result of accuracy of several algorithms including Naïve Bayes, Nearest Neighbor, Support Vector Machine (SVM), Decision Tree, and Multilayer Perceptron Neural Network. The result shows that SVM outperform by achieving 100% of accuracy. According to this result, it concludes the excellent performance of SVM in handling high dimensional of data.

Keywords – cyber terrorism, data mining, feature selection, text classification, web mining

DEDICATION

I dedicate this thesis to my family who has supported me all the way.



ACKNOWLEDGMENTS

The author wishes to give utmost gratitude to Jesus Christ, without His love and blessing all of the work and effort will be nothing, praise and glory only for His name.

The author would also give the best regard to Anto Satriyo Nugroho, Dr.Eng and Charles Lim, Msc for the guidance, encouragement, patience, and their assistance from the beginning until the end of this thesis research. So many thanks also addressed to all of the lecturers and staffs of Swiss German University.

Special regards also comes for my fellow students in Information Technology 2006, I gratitude you guys for these beautiful four years in my life, especially Endy Chen for being such a collaborative partner in conducting this research.

In addition, the author would also like to give thanks to Armando, Irfan Suwadi, Ivan Firdausi, and Ryan Andhika Perdana for being such a mutual friends and very helpful to me.

Last but not least, the author would give special thanks to Alicia Item for being my perfect distraction but also my spirit booster.

TABLE OF CONTENTS

| | |
|--|----|
| STATEMENT BY THE AUTHOR..... | 2 |
| ABSTRACT..... | 3 |
| DEDICATION..... | 4 |
| ACKNOWLEDGMENTS..... | 5 |
| 1 CHAPTER 1 – INTRODUCTION..... | 10 |
| 1.1 Background..... | 10 |
| 1.2 General Problem Statement..... | 11 |
| 1.3 Research Objectives..... | 11 |
| 1.4 Significance of Study..... | 11 |
| 1.5 Research Scope..... | 11 |
| 1.6 Methodology..... | 13 |
| 2 CHAPTER 2 – LITERATURE REVIEW..... | 14 |
| 2.1 Cyber Terrorism..... | 14 |
| 2.1.1 Definition..... | 14 |
| 2.1.2 Elements in Cyber Terrorism..... | 15 |
| 2.1.3 Objectives and Types of Attack in Cyber Terrorism..... | 17 |
| 2.2 Web Mining..... | 18 |
| 2.2.1 Web Structure Mining..... | 19 |
| 2.2.2 Web Usage Mining..... | 19 |
| 2.2.3 Web Content Mining..... | 19 |
| 2.3 Text Classification..... | 19 |
| 2.3.1 Classifiers..... | 20 |
| 2.3.2 Naïve Bayes..... | 21 |
| 2.3.3 Decision Tree C4.5..... | 22 |
| 2.3.4 Nearest Neighbor..... | 23 |
| 2.3.5 Multilayer Perceptron Neural Network..... | 25 |
| 2.3.6 Support Vector Machine..... | 29 |
| 2.3.6.1 Linear SVM..... | 29 |
| 2.3.6.2 Soft Margin..... | 31 |
| 2.3.6.3 Non-Linear SVM..... | 32 |
| 3 CHAPTER 3 – METHODOLOGY..... | 35 |
| 3.1 Dataset..... | 35 |
| 3.1.1 Holdout Method..... | 35 |
| 3.2 General Idea..... | 36 |
| 3.3 Data Preprocessing..... | 37 |
| 3.3.1 Tokenization..... | 37 |
| 3.3.2 Feature Selection..... | 39 |
| 3.3.2.1 Best-First Algorithm..... | 40 |
| 3.3.3 Vector Generation..... | 41 |
| 3.3.4 Attribute Relation File Format (ARFF) Files..... | 44 |
| 4 CHAPTER 4 – RESULT & DISCUSSION..... | 46 |
| 4.1 Result..... | 46 |
| 4.1.1 K-Nearest Neighbor..... | 47 |

| | | |
|---------|--|-----|
| 4.1.2 | Naïve Bayes | 51 |
| 4.1.3 | Support Vector Machine (SVM)..... | 51 |
| 4.1.4 | Decision Tree C4.5 | 53 |
| 4.1.5 | Multilayer Perceptron | 54 |
| 4.1.6 | Performance Metrics | 54 |
| 4.1.6.1 | Without Feature Selection | 57 |
| 4.1.6.2 | Feature Selection | 60 |
| 4.1.6.3 | Performance Comparison Between with and without Feature Selection | 63 |
| 4.2 | Discussion | 65 |
| 4.2.1 | Instance Number 52 | 66 |
| 5 | CHAPTER 5 – CONCLUSION AND RECOMMENDATION | 69 |
| 5.1 | Conclusion..... | 69 |
| 5.2 | Future Works..... | 70 |
| 6 | GLOSSARY | 71 |
| 7 | REFERENCES | 74 |
| 8 | PUBLICATION..... | 77 |
| 9 | APPENDICES | 78 |
| | Appendix A. Training Set References | 78 |
| | Appendix B. Test Set References | 98 |
| | Appendix C. Dictionary | 108 |
| | Appendix D. Instance Number 52 | 117 |
| | Appendix E. Instance Number 10..... | 120 |
| | Appendix F. Publication | 128 |
| 10 | CURRICULUM VITAE..... | 134 |



SWISS GERMAN UNIVERSITY