

Universiti Teknologi MARA

**Text Categorization with Support Vector
Machines**

Nur Amira Binti Khanafi

Thesis submitted in fulfillment of the requirements for

Bachelor of Science (Hons) Intelligent System

Faculty of Computer and Mathematical Sciences

July 2013

ACKNOWLEDGMENT

In The Name Of Allah Most Gracious Most Merciful

Alhamdulillah, Praise and thanks to Allah for His blessing and granted the courage and patience to complete my final year project. First and foremost, I would like to express my deepest sincere and gratitude to my supervisor, Dr Shuzlina Abdul Rahman for her continuous guidance, understanding, encouragement and valuable advices. I also would like to express my gratitude to Dr Marina Yusoff, Final Year Project Coordinator and also all lecturers for the help in completing the project. Thanks also to my dearest friend Yasmin Mohamad Shafiee, Nadirah Hanisah Abu Bakar and my fellow course mates for helping me in completing my project. Lastly, thanks to my family, especially for my lovely parents Khanafi B. Hj. Atan and [REDACTED] that gave encouragement and supporting me in term of care and love until I was able to complete my project effectively. Thank you. May Allah bless us all.

ABSTRACT

With the growing number of text documents on the internet made it difficult for users to search, find, manage and organize information quickly. In the past, text documents are classified manually and it is time-consuming. Text categorization is a process of assigning text documents into a set of fixed predefined categories. The high dimensionality of text documents made it difficult to categorize because text documents contain noise and useless data. This project explained several methods of feature selection that can be used to reduce high dimensionality of feature space in text documents such as Information Gain, Gain Ratio, CHI-Squares, Mutual Information and Document frequency. This project also explained how text categorization is developed using Support Vector Machines and solved the text categorization problems. The results showed that Support Vector Machines perform well and very fast at both training and testing. It also showed that Support Vector Machines is a suitable technique to handle text categorization since this project are using large amount of datasets.

TABLE OF CONTENTS

CONTENTS	PAGE
SUPERVISOR'S APPROVAL	i
DECLARATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	viii
LIST OF TABLES	ix
LIST OF EQUATIONS	x

CHAPTER ONE: INTRODUCTION

1.1	Research Background	1
1.2	Problem Statement	2
1.3	Research Questions	2
1.4	Research Objective	3
1.5	Research Scope	3
1.6	Research Significance	3
1.7	Project Overview	4
1.8	Thesis Outline	5

CHAPTER TWO: LITERATURE REVIEW

2.1	Text Categorization (TC)	7
2.2	Methods of Pre-Processing for Text Documents	8

CHAPTER 1

INTRODUCTION

This chapter explains about the introduction and some brief explanation of the project. The main components of the research include the research background, the problem statement, the research questions and the research objective, the project scope and the research significance. This chapter ends by describing the outline of the thesis.

1.1 Research Background

With the rapid growth of the internet, web documents that contain a lot of textual information have made it difficult for users to searching information quickly on the web. To classify millions of text documents manually is very difficult and taking up much time (Alsaleem, 2011). So, this led to developing software or tools that can help users searching information which is the accuracy and time efficiency is much better than the manual classification. The main issue in the text categorization is the process of classifying the text documents into fixed pre-defined categories. Text categorization is become one of the most important and powerful components in many organizations and on the internet for organizing and managing task or documents. Furthermore, it has ability to classifying many types of task or documents. It also helps the users in finding the data and related information.