

## ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to those who helped me with various aspects of conducting research and writing this thesis.

Firstly, I would like to express my deepest gratitude to **Dr. Mie Mie Thet Thwin**, Rector, the University of Computer Studies, Yangon, for her kind permission to develop this thesis.

I am grateful to my thesis supervisor **Dr. Khin Mar Soe**, Professor, NLP Lab, for her suggestions and encouragement to do this thesis and also for her close supervision, invaluable suggestions, kind guidance and constant encouragement during the course of this work.

I also like to express my thanks to our dean **Dr. Thi Thi Soe Nyunt**, Professor and Head of Faculty of Computer Science, for her kind help in this work.

I would like to extend my deepest gratitude to **Daw Ni Ni San**, Lecturer, English Department, for her advice and language editing.

I heartily appreciate the suggestions and recommendations of the teachers who attended all my seminars.

Sincere thanks are also due to my close friends for their kind help, understanding and cooperation throughout the work. Finally, I am most indebted to my beloved family for their support, patience and constant encouragement throughout my studies.

## **ABSTRACT**

Text Classification is the task of automatically assigning a set of documents into certain categories (class or topics) from a predefined set. This also plays an important role in natural language processing and also crossroads between information retrieval and machine learning. The dramatic growth of text document in digital form news website makes the task of text classification more popular over last ten year. The application of this method can be found in spam filtering, question and answering, language identification. This paper presents the idea of text classification process in term of using machine learning technique and illustrates how Myanmar news documents are classified by applying genetic algorithm. The applied system will use Myanmar online news articles from Myanmar news website for the purpose of training and testing the system. Term frequency inverse document frequency (TF-IDF) algorithm was used to select related feature according to their labelled document which is also applied in many text mining methods.

# TABLE OF CONTENTS

	<b>Page</b>
<b>ACKNOWLEDGEMENTS</b>	i
<b>ABSTRACT</b>	ii
<b>TABLE OF CONTENTS</b>	iii
<b>LIST OF FIGURE</b>	vi
<b>LIST OF TABLES</b>	viii
<b>LIST OF EQUATIONS</b>	ix
<b>CHAPTER 1 INTRODUCTION</b>	
1.1 Objective of Thesis	2
1.2 Overview of Thesis	2
1.3 Organization of Thesis	2
<b>CHAPTER 2 THEORETICAL BACKGROUND</b>	
2.1 Text Mining	3
2.1.1 Text Mining and Its Applications	3
2.2 Text Classification	4
2.2.1 The importance of Automatic Text Classification	5
2.3 Text Classification Operations	5
2.3.1 Text Data Preprocessing	7
2.3.1.1 Stop Words Removal	7
2.3.1.2 Word Segmentation	7
2.3.1.3 Feature Selection	8
(a) Gini Index	8

	(b) Information Gain	8
	(c) Mutual Information	9
	(d) Chi-Square	9
	(e) TF-IDF	9
2.3.2	Text Classification	10
2.3.3	Supervised Learning	12
	(a) Decision Tree	13
	(b) Native Bayesian Classifier	13
	(c) Support Vector Machine	14
	(d) K-Nearest Neighbor	15
2.3.4	Unsupervised Learning	15
	(a) Genetic Algorithm	16
	(b) Neural Network	17

## **CHAPTER 3 TEXT CLASSIFIER SYSTEM FOR MYANMAR NEWS ARTICLES**

3.1	Text Mining and Text Classifier	18
3.2	Myanmar Language Nature	19
	3.2.1 Preprocessing in Myanmar Language	20
3.3	Feature Selection for Text Classification	22
3.4	Genetic Algorithm	23
	(1) Initial Population	23
	(2) Fitness Function	24
	(3) Selection	25

	(4) Crossover	25
	(5) Mutation	26
	(6) Termination	26
3.5	Algorithm of Text Classification System for Myanmar News Articles	26
3.6	A Working Example	28
	3.6.1 Dataset	27
	3.6.2 Training Dataset	27
	3.6.3 Classification of Documents	33
<b>CHAPTER 4</b>	<b>DESIGN AND IMPLEMENTATION OF THE SYSTEM</b>	
4.1	Design of the System	46
4.2	Implementation of the System	47
	4.2.1 Data Collection	47
	4.2.2 Preparation for Training Dataset	48
	4.2.3 User Interfaces of the System	49
4.3	Experimental Result	56
<b>CHAPTER 5</b>	<b>CONCLUSION</b>	
5.1	Advantages and Limitation of the System	57
5.2	Application Area	58
<b>REFERENCES</b>		59

## **LIST OF FIGURES**

<b>Figure</b>		<b>Page</b>
Figure	3.1 A sample sentence from document	21
Figure	3.2 Segmented Sample Sentence	21
Figure	3.3 Stopword Removal of Sample Sentence	22
Figure	3.4 Population, Chromosomes and Genes	25
Figure	3.5 Proposed Algorithm for Myanmar Text Classifier using Genetic algorithm	26
Figure	3.6 Politic Document	27
Figure	3.7 Segmentation of Sample Document	28
Figure	3.8 Removing Stopword from Sample Document	28
Figure	3.9 Collected Terms from Sample Document	29
Figure	3.10 Input Document	33
Figure	3.11 Segmentation of Input Document	34
Figure	3.12 Removal of Stopword form Input Document	35
Figure	3.13 Algorithm for Tournament Selection	41
Figure	4.1 Class diagram for Myanmar Text Classifier using Genetic Algorithm	47
Figure	4.3 Main Page of the System	53
Figure	4.4 Training page of the System	53
Figure	4.5 Training Document	54
Figure	4.6 Page view after training documents.	54
Figure	4.7 Classification Page	55
Figure	4.8 Classifying Documents	56
Figure	4.9 Initial Stage of the classification	56

Figure	4.10	Stop Word Removal Stage	57
Figure	4.11	Feature Extraction Stage	58
Figure	4.12	Category of the input document	58

## LIST OF TABLES

<b>Table</b>		<b>Page</b>
Table 3.1	Calculation of TF-IDF	30
Table 3.2	Collected Feature Words	33
Table 3.3	Collected Term from the Input Document	36
Table 3.4	Term and Chromosome	37
Table 3.5	Tem with each frequency and chromosome	39
Table 3.6	Fitness Value for Each Term	40
Table 3.7	Child Population	43
Table 3.8	2 <sup>nd</sup> Generation Child Chromosome	44
Table 3.9	3 <sup>rd</sup> Generation Child Chromosome	45
Table 4.1	Document Collection for Training and Test Data	48
Table 4.2	Stopword List	49
Table 4.3	Accuracy Measurement of the System	56



## LIST OF EQUATIONS

<b>Equation</b>			<b>Page</b>
Equation	2.1	Equation for Information Gain	8
Equation	2.2	Mutual Information	9
Equation	2.3	Calculation of the Global Goodness of a term	9
Equation	2.4	Calculation of the Global Goodness of a term	9
Equation	2.5	Calculation of Chi-Square	9
Equation	2.6	Calculation of Term Frequency	10
Equation	2.7	Calculation of Inverse Document Frequency	10
Equation	2.8	Calculation of TF-IDF	10
Equation	3.1	Calculation of Weight Term Standard Deviation	24
Equation	3.2	Calculation of Term Frequency	29
Equation	3.3	Calculation of Inverse Document Frequency	29
Equation	3.4	Calculation of TF-IDF	29
Equation	3.5	Calculation of Weight Term Standard Deviation	39
Equation	4.1	Calculation of Precision	56

Equation	4.2	Calculation of Recall	56
Equation	4.3	Calculation of F1 Score	56