

## TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	<b>ABSTRACT</b>	iii
	<b>LIST OF TABLES</b>	xii
	<b>LIST OF FIGURES</b>	xv
	<b>LIST OF SYMBOLS AND ABBREVIATIONS</b>	xvii
1	<b>INTRODUCTION</b>	1
	1.1 SENTIMENT ANALYSIS	1
	1.2 SENTIMENT ANALYSIS METHODS	1
	1.2.1 Data Mining Approach	2
	1.2.2 Natural Language Processing Approach	3
	1.3 MACHINE LEARNING ALGORITHMS	4
	1.3.1 The General Structure of the Problem	5
	1.3.2 Probabilistic Classifiers –The Naive Bayes Algorithm	5
	1.3.3 Bayesian Logistic Regression	5
	1.3.4 Decision Tree Classifiers	6
	1.3.5 Neural Networks	6
	1.3.6 Regression Methods	7
	1.3.7 Support Vector Machines	7
	1.4 FUZZY LOGIC	8
	1.5 MAHALANOBIS DISTANCE BASED CLASSIFICATION	8
	1.5.1 Mahalanobis Distance	9
	1.6 ORGANIZATION OF THIS THESIS	9

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
<b>2</b>	<b>LITERATURE REVIEW</b>	11
2.1	SENTIMENT ANALYSIS	11
2.2	SENTIMENT ANALYSIS USING MACHINE LEARNING ALGORITHMS	11
2.3	SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING APPROACH	15
2.4	SENTIMENT ANALYSIS USING HYBRID APPROACHES	30
2.5	SENTIMENT ANALYSIS USING FUZZY LOGIC	31
2.6	MAHALANOBIS DISTANCE FOR TEXT CATEGORIZATION	34
2.7	SCOPE AND OBJECTIVE	34
	2.7.1 Scope	34
	2.7.2 Objective	35
2.8	METHODOLOGY	35
	2.8.1 The Flow of Activities in this Thesis	36
<b>3</b>	<b>DEVELOPMENT OF REPRESENTATIVE TERM-DOCUMENT MATRIX</b>	37
3.1	INTRODUCTION	37
3.2	REPRESENTATIVE TERM-DOCUMENT MATRIX (RTDM) CREATION	38
3.3	DATASET USED IN THIS THESIS	42
3.4	PERFORMANCE MEASURES	43
3.5	CLASSIFICATION PERFORMANCE OF VARIOUS CLASSIFIERS	44

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
3.6	DISCUSSION ON THE CLASSIFICATION PERFORMANCE OF ML CLASSIFIERS WITH RTDM AS INPUT	45
3.7	ASSUMPTIONS AND SOME SPECIAL OBSERVATIONS	46
<b>4</b>	<b>DESIGN AND DEVELOPMENT OF MAHALANOBIS DISTANCE BASED CLASSIFIER (MDC)</b>	<b>47</b>
4.1	INTRODUCTION	47
4.2	MD FOR PATTERN RECOGNITION	47
4.2.1	Calculation of MD	48
4.3	SENTIMENT DETECTION USING MD	49
4.3.1	RTDM Creation from the Review Documents	49
4.3.2	Selection of Mahalanobis Space (MS)	49
4.3.3	Calculation of MD	51
4.3.4	Determination of “Threshold MD” for Sentiment Classification	51
4.3.5	Sentiment Classification	52
4.4	RESULTS	52
4.5	DISCUSSIONS	58
4.6	IMPORTANT OBSERVATIONS	59
<b>5</b>	<b>DESIGN AND DEVELOPMENT OF FUZZY LOGIC CLASSIFIER (FLC) FOR SENTIMENT ANALYSIS</b>	<b>61</b>
5.1	INTRODUCTION	61
5.2	FUZZY CLASSIFIER	62
5.2.1	Fuzzy Set	62

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
	5.2.2 Linguistic Variables and Hedges	63
5.3	GENERAL STEPS INVOLVED IN THE FUZZY INFERENCE	64
5.4	SENTIMENT CLASSIFICATION USING FUZZY LOGIC	65
	5.4.1 Introduction	65
	5.4.2 Linguistic Variables for the Sentiment Analysis Process	66
	5.4.3 RTDM Creation	66
	5.4.4 Creation of RTDM with Fuzzy Scores (RTDM-FS)	66
	5.4.5 Fuzzy Score (FS) Calculation	67
5.5	RESULTS AND DISCUSSION	72
<b>6</b>	<b>DESIGN AND DEVELOPMENT OF MAHALANOBIS- FUZZY SYSTEM CLASSIFIER (MFSC) FOR SENTIMENT ANALYSIS</b>	<b>75</b>
	6.1 INTRODUCTION	75
	6.2 SENTIMENT CLASSIFICATION USING MFSC	75
	6.3 RESULTS	78
	6.4 DISCUSSIONS AND IMPORTANT OBSERVATIONS	84
<b>7</b>	<b>DESIGN AND DEVELOPMENT OF HYBRID CLASSIFIERS FOR SENTIMENT ANALYSIS</b>	<b>85</b>
	7.1 INTRODUCTION	85
	7.2 HYBRID OF MDC AND ML BASED CLASSIFIERS	86

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
7.3	HYBRID OF MFSC AND ML BASED CLASSIFIERS	87
7.4	HYBRID OF FLC AND ML BASED CLASSIFIERS	89
7.5	DISCUSSION AND INFERENCE	91
<b>8</b>	<b>PERFORMANCE COMPARISON OF MDC, MFSC AND FLC</b>	<b>92</b>
8.1	INTRODUCTION	92
8.2	COMPARISON OF PERFORMANCE (FOR CAMERA AND CELL PHONE REVIEWS)	93
8.3	COMPARISON OF PERFORMANCE (FOR MOVIE REVIEWS - LDS)	96
<b>9</b>	<b>CONCLUSION</b>	<b>100</b>
9.1	CONCLUSION	100
9.2	SCOPE FOR FUTURE RESEARCH	101
	<b>APPENDIX 1: RTDM OF LDS 403 WITH MD VALUES</b>	<b>103</b>
	<b>REFERENCES</b>	<b>114</b>
	<b>LIST OF PUBLICATIONS</b>	<b>124</b>
	<b>CURRICULUM VITAE</b>	<b>125</b>

## LIST OF TABLES

TABLE NO.	TITLE	PAGE NO.
3.1	List of stop words	40
3.2	A sample Representative Term-Document Matrix (RTDM)	42
3.3	Performance of various classifiers for LDS403 (202 negative and 201 positive)	44
3.4	Performance of various classifiers for LDS2000 (1000 negative and 1000 positive)	44
3.5	Performance of various classifiers for LDS11000 (5500 negative and 5500 positive)	45
3.6	Performance of various classifiers for LDS25000 (12500 negative and 12500 positive)	45
4.1	The MS chosen for analysis of movie reviews	50
4.2	Abridged RTDM of LDS403	50
4.3	Abridged RTDM of LDS403 with MD of each review	51
4.4	Confusion matrix for camera reviews (MDC)	53
4.5	Confusion matrix for cell phone reviews (MDC)	53
4.6	Confusion matrix for LDS403 (MDC)	53
4.7	Confusion matrix for LDS2000 (MDC)	53
4.8	Confusion matrix for LDS11000 (MDC)	53
4.9	Confusion matrix for LDS25000 (MDC)	54
4.10	Classification performance of MDC on various datasets	54
4.11	Details of MS for various datasets	58

<b>TABLE NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
5.1	Procedure for calculation of fuzzy weight of linguistic variables	64
5.2	Fuzzy weight of linguistic variables	67
5.3	Feature vector with fuzzy score for the review shown in Figure 5.1	70
5.4	Abridged RTDM-FS for LDS403 with fuzzy score and assigned class label	71
5.5	Confusion matrix for camera reviews (FLC)	72
5.6	Confusion matrix for cell phone reviews (FLC)	72
5.7	Confusion matrix for LDS403 (FLC)	72
5.8	Confusion matrix for LDS2000 (FLC)	73
5.9	Confusion matrix for LDS11000 (FLC)	73
5.10	Confusion matrix for LDS25000 (FLC)	73
5.11	Classification performance of FLC on various datasets	73
6.1	Abridged RTDM-FS for LDS403	76
6.2	MS for the Mahalanobis-Fuzzy System Classifier	77
6.3	Abridged RTDM-FS with MD score for LDS403 dataset	77
6.4	Confusion matrix for camera reviews (MFSC)	78
6.5	Confusion matrix for cell phone reviews (MFSC)	79
6.6	Confusion matrix for LDS403 (MFSC)	79
6.7	Confusion matrix for LDS2000 (MFSC)	79
6.8	Confusion matrix for LDS11000 (MFSC)	79
6.9	Confusion matrix for LDS25000 (MFSC)	79
6.10	Classification performance of MFSC on various datasets	80
6.11	Details of MS for various datasets	84

<b>TABLE NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
7.1	Performance of various ML classifiers on the documents misclassified by MDC	87
7.2	Performance of various ML classifiers on the documents misclassified by MFSC	88
7.3	Performance of various ML classifiers on the documents misclassified by FLC	90
7.4	Best performing hybrid classifiers (for LDS25000)	91



## LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
2.1	The flow of activities in this study	36
3.1	A sample review and the procedure for creating feature vector using RT	41
4.1	MD plot for camera reviews	55
4.2	MD plot for cell phone reviews	55
4.3	MD plot for LDS403	56
4.4	MD plot for LDS2000	56
4.5	MD plot for LDS11000	57
4.6	MD plot for LDS25000	57
5.1	A sample review and its feature vector	68
6.1	MD plot for camera reviews (MFSC)	81
6.2	MD plot for cell phone reviews (MFSC)	81
6.3	MD plot for LDS403 (MFSC)	82
6.4	MD plot for LDS2000 (MFSC)	82
6.5	MD plot for LDS11000 (MFSC)	83
6.6	MD plot for LDS25000 (MFSC)	83
7.1	The hybrid classification experimental procedure for MDC	86
7.2	The hybrid classification experimental procedure for MFSC	88
7.3	The hybrid classification experimental procedure for FLC	89

<b>FIGURE NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
8.1	Comparison of Precision – MDC, MFSC and FLC (Camera and Cell Phone Reviews)	93
8.2	Comparison of Recall – MDC, MFSC and FLC (Camera and Cell Phone Reviews)	94
8.3	Comparison of F-Measure – MDC, MFSC and FLC (Camera and Cell Phone Reviews)	95
8.4	Comparison of Accuracy – MDC, MFSC and FLC (Camera and Cell Phone Reviews)	95
8.5	Comparison of Precision – MDC, MFSC and FLC (for the subsets of LDS)	96
8.6	Comparison of Recall – MDC, MFSC and FLC (for the subsets of LDS)	97
8.7	Comparison of F-Measure – MDC, MFSC and FLC (for the subsets of LDS)	98
8.8	Comparison of Accuracy – MDC, MFSC and FLC (for the subsets of LDS)	99

## LIST OF SYMBOLS AND ABBREVIATIONS

A	- Accuracy
ANN	- Artificial Neural Networks
ASUM	- Aspect and Sentiment Unification Model
BLR	- Bayesian Logistic Regression
CCA	- Canonical Correlation Analysis
c	- Category
CSR	- Class Sequential Rule mining
CART	- Classification And Regression Tree
C	- Comparative words
CRF	- Conditional Random Fields
CDTI	- Cross Domain Topic Indexing
DT	- Decision Tree
d	- Document
FN	- False Negative
FP	- False Positive
F	- Features
FLC	- Fuzzy Logic Classifier
FOM	- Fuzzy Opinion Miner
FS	- Fuzzy Score
FW	- Fuzzy Weight
GAMS	- General Algebraic Modeling. System
GIBC	- General Inquirer Based Classifier
GSP	- Gram-Schmidt Process
HMM	- Hidden Markov Model

IR	-	Information Retrieval
IMDB	-	Internet Movie DataBase
$C^{-1}$	-	Inverse of the Correlation Matrix of $Z_{ij}$
KNN	-	K Nearest Neighbor
L-EM	-	Labeled Expectation Maximization
LDS	-	Large movie review DataSet
LDA	-	Latent Dirichlet Allocation
LLSF	-	Linear Least Square Fit
$\varphi(x)$	-	Logistic link function
LR	-	Lucene Rank
ML	-	Machine Learning
MD	-	Mahalanobis Distance
MDC	-	Mahalanobis Distance based Classifier
MFS	-	Mahalanobis Fuzzy System
MFSC	-	Mahalanobis Fuzzy System Classifier
MS	-	Mahalanobis Space
MTS	-	Mahalanobis Taguchi System
$P(d)$	-	Marginal probability
MAX	-	Maximum
ME	-	Maximum Entropy
MAXOC	-	Maximum number of OCcurrences
MST	-	Maximum Spanning Trees
MIN	-	Minimum
MINOC	-	Minimum number of OCcurrences
$\beta$	-	Model parameters vector
MLP	-	Multi Layer Perceptron
NB	-	Naïve Bayes
NBm	-	Naïve Bayes multinomial
NLP	-	Natural Language Processing

NN	- Neural Network
K	- Number of features or variables
OSA	One-Side Association
OM	- Opinion Mining
OSR	- Overall Success Rate
POS	- Parts Of Speech
PMI	- Point-wise Mutual Information
PSP	- Positive Sentence Percentage
PERL	- Practical Extraction and Report Language
P	- Precision
PCS	- Preferred entities in Comparative Sentences
PSLA	- Probabilistic Latent Semantic Analysis
$P(c d)$	- Probability that the document $d$ belongs to the category $c$
R	- Recall
RT	- Representative term
RTDM	- Representative Term-Document Matrix
RT	- Representative Terms
RBC	- Rule Based Classifiers
SO	- Semantic Orientation
SO-CAL	- Semantic Orientation CALculator
SLDA	- Sentence Latent Dirichlet Allocation
SA	- Sentiment Analysis
SMO	- Sequential Minimal Optimization
D	- Set of all Documents
SVD	- Singular Value Decomposition
SFA	- Spectral Feature Alignment
s	- Standard Deviation
$Z_{ij}$	- Standardized matrix
SAS	- Statistical Analysis System

SBC	-	Statistics Based Classifier
SVM	-	Support Vector Machine
TOQ	-	Temporal Opinion Quality
TDM	-	Term Document Matrix
TC	-	Text Categorization
TTDM	-	Topical Term Description Model
TF	-	Total Frequency
$Z_{ij}^T$	-	Transpose of $Z_{ij}$
TN	-	True Negative
TP	-	True Positive
VEO	-	Vector of Emotional Occurrences
WEKA	-	Waikato Environment for Knowledge Analysis