

## TABLE OF CONTENT

UNDERGRADUATE THESIS .....	i
BACHELOR'S THESIS .....	i
APPROVAL PAGE .....	ii
APPROVAL PAGE .....	iii
STATEMENT OF ORIGINALITY .....	iv
TABLE OF CONTENT .....	v
LIST OF TABLES .....	vii
LIST OF FIGURES .....	viii
ABSTRACT .....	ix
INTISARI .....	x
CHAPTER 1 INTRODUCTION .....	1
1.1 Research Background .....	1
1.2 Research Problem .....	3
1.3 Research Scope .....	3
1.4 Research Objective .....	3
1.5 Research Benefit .....	4
1.6 Research Steps .....	4
CHAPTER 2 LITERATURE REVIEW .....	5
Chapter 3 THEORETICAL BASIS .....	8
3.1 Text Mining .....	8
3.2 Preprocessing .....	9
3.3 Topic Modeling-LDA algorithm .....	11
3.3.1 Topic Modeling .....	11
3.3.2 Latent Dirichlet Allocation (LDA) .....	12
3.3.3 Topic Coherence of LDA .....	13
3.4 Sentiment Analysis .....	14
3.5 TF-IDF .....	15
3.6 Naïve Bayes .....	16
3.7 K-Nearest Neighbor .....	18
3.8 Confusion matrix Evaluation .....	18
3.9 Culinary Target Marketing .....	19
CHAPTER 4 RESEARCH METHODOLOGY .....	21
4.1 General Analysis .....	21
4.2 Data Acquisition .....	23
4.3 Preprocessing .....	23
4.3.1 Transform case .....	23
4.3.2 Tokenization .....	24
4.3.3 Stopwords Removal .....	24
4.3.4 Stemming .....	24

4.4	Experiment .....	25
4.4.1	Topic modeling-LDA .....	25
4.4.2	Evaluation of LDA .....	26
4.4.3	Sentiment Analysis .....	26
4.4.4	K Nearest Neighbor .....	26
4.4.5	Multinomial Naïve Bayes .....	27
CHAPTER 5	IMPLEMENTATION .....	28
5.1	Specification.....	28
5.2	Data Pre-processing Implementation .....	28
5.2.1	Data Deduplication and tags removal.....	32
5.2.2	Stopwords .....	32
5.2.3	Tokenization and Transform case.....	33
5.2.4	Stemming.....	34
5.3	Topic Modeling.....	35
5.3.1	Determine the number iteration and topics.....	37
5.4	Coherence Value Evaluation.....	38
5.5	Sentiment Analysis .....	39
5.5.1	Training and Testing.....	41
5.5.2	Multinomial Naïve Bayes .....	41
5.5.3	Confusion matrix of Naïve Bayes .....	42
5.5.4	K-Nearest Neighbors Classification .....	43
CHAPTER 6	RESULTS AND DISCUSSION.....	47
6.1	Data Acquisition result.....	47
6.2	Data Pre-processing result .....	48
6.3	Topic modeling result .....	48
6.3.1	LDA Result.....	49
6.3.2	LDA Visualization.....	50
6.3.3	LDA Evaluation.....	51
6.4	Sentiment analysis result.....	52
6.4.1	Multinomial Naïve Bayes result .....	52
6.4.2	K-Nearest Neighbor result.....	53
CHAPTER 7	CONCLUSION AND FUTURE WORKS.....	56
7.1	Conclusions.....	56
7.2	Limitation.....	56
7.3	Future works .....	56
References	.....	57

## **LIST OF TABLES**

Table 2.1	Comparison with previous research .....	7
Table 3.1	Examples of Snowball algorithm .....	10
Table 4.1	Example of Transform case .....	24
Table 4.2	Example of Tokenization.....	24
Table 4.3	Example of Stopwords .....	24
Table 4.4	Example of Stemming .....	25
Table 6.1	Data size before and after Preprocess.....	48
Table 6.2	Words in corpus 6 and 7 topics .....	49
Table 6.3	Coherence score throughout the range (1-49) .....	51

## LIST OF FIGURES

Figure 3.1	Text Mining Process in Text Document Processing .....	8
Figure 3.2	Concept of topic modeling (Blei, 2012) .....	11
Figure 3.3	Topics from articles in the New York Times (Hoffman et al., 2013) .....	12
Figure 3.4	Graphical model representation of LDA .....	13
Figure 3.5	The TP, TN, FP, and FN values in confusion matrix. ....	19
Figure 3.6	Example of culinary trend analysis .....	20
Figure 4.1	Research design diagram .....	22
Figure 4.2	Preprocessing Steps .....	23
Figure 4.3	Steps of LDA .....	26
Figure 4.4	Steps of KNN .....	27
Figure 4.5	Steps of Multinomial Naïve bayes.....	27
Figure 5.1	Code for preprocessing phase.....	32
Figure 5.2	Code for deduplication and tags removal .....	32
Figure 5.3	Code for Stopwords and extra Stopwords .....	33
Figure 5.4	Code for Tokenization and transform case.....	34
Figure 5.5	Code to run Stemming .....	34
Figure 5.6	Code for the topic modeling .....	37
Figure 5.7	Code to plot the coherence and perplexity value.....	38
Figure 5.8	Printing the LDA model-pyLDAvis .....	38
Figure 5.9	Code to calculate coherence score throughout the range (1-49) .....	39
Figure 5.10	Sentiment analysis code .....	41
Figure 5.11	Training and Testing.....	41
Figure 5.12	Implementation of Multinomial Naïve Bayes .....	42
Figure 5.13	Code for Naïve Bayes confusion matrix.....	43
Figure 5.14	Implementation of KNN .....	46
Figure 6.1	Showing example of the first 5 data from Amazon fine food review .....	47
Figure 6.2	Showing the coherence score throughout the range of topics .....	49
Figure 6.3	LDA with 6 topic .....	50
Figure 6.4	LDA with 7 topic .....	50
Figure 6.5	ROC Curve of multinomial Naïve Bayes .....	52
Figure 6.6	Naive Bayes confusion matrix.....	53
Figure 6.7	GridSearch cross validation result.....	54
Figure 6.8	ROC curve for KNN.....	54
Figure 6.9	KNN confusion matrix .....	55